TEMPLATE MATCHING TECHNIQUES IN COMPUTER VISION

THEORY AND PRACTICE

Roberto Brunelli

Fondazione Bruno Kessler, Italy

WILEY

A John Wiley and Sons, Ltd, Publication

TEMPLATE MATCHING TECHNIQUES IN COMPUTER VISION

TEMPLATE MATCHING TECHNIQUES IN COMPUTER VISION

THEORY AND PRACTICE

Roberto Brunelli

Fondazione Bruno Kessler, Italy

WILEY

A John Wiley and Sons, Ltd, Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Brunelli, Roberto, 1961-
Template matching techniques in computer vision : theory and practice / Roberto Brunelli.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-51706-2 (cloth)
1. Computer vision. 2. Template matching (Digital image processing) 3. Image processing–Digital techniques. 4. Optical pattern recognition. 5. Image analysis. I. Title.
TA1634.B78 2009
006.3'7–dc22
2008052237

A catalogue record for this book is available from the British Library.

ISBN 978-0-470-51706-2

Set in 10/12pt Times by Sunrise Setting Ltd, Torquay, UK.
Printed in Great Britain by CPI Antony Rowe, Chippenham, Wiltshire.

CONTENTS

Appendices 293

PREFACE

Detection and recognition of objects from their images, irrespective of their orientation, scale, and view, is a very important research subject in computer vision, if not computer vision itself. This book focuses on a subset of the object recognition techniques proposed so far by the computer vision community, those employing the idea of projection to match image patterns, and on a specific class of target objects, faces, to illustrate general object recognition approaches. Face recognition and interpretation is a critical task for people, and one at which they excel. Over the last two decades it has received increasing attention from the computer vision community for a variety of reasons, ranging from the possibility of creating computational models of interesting human recognition tasks, to the development of practical biometric systems and interactive, emotion-aware, and capable human–machine interfaces.

The topics covered in this book have been investigated over a period of about 30 years by the image processing community, providing increasingly better computer vision solutions to the problem of automatic object location and recognition. While many books on computer vision are currently available that touch upon some of the topics addressed in the present book, none of them, to the best of the author's knowledge, provides a coherent, in-depth coverage of template matching, presenting a varied set of techniques from a common perspective. The methods considered present both theoretical and practical interest by themselves as well as enabling techniques for more complex vision systems (stereo vision, robot navigation, image registration, multimedia retrieval, target tracking, landmark detection, just to mention a few). The book contains many photographs and diagrams that help the user grasp qualitative and quantitative aspects of the material presented. The software available on the book's web site provides a high-level image processing environment and image datasets to explore the techniques presented in the book.

Knowledge of basic calculus, statistics, and probability theory is a prerequisite for the reader. The material covered in the book is at the level of (advanced) undergraduate students or introductory Ph.D. courses and will prove useful to researchers and developers of computer vision systems addressing a variety of tasks, from robotic vision to quality control and biometric systems. It may be used for a special topics course on image analysis at the graduate level. Another expected use is as a supporting textbook for an intensive short course on template matching, with the possibility of choosing between a theoretical and an application-oriented bias. The techniques are discussed at a level that makes them useful also for the experienced researcher and make the book an essential learning kit for practitioners in academia and industry.

Rarely, if ever, does a book owe its existence to the sole author, and this one certainly does not. First a tribute to the open source software community, for providing the many tools necessary to describe ideas and making them operational. To Jaime Vives Piqueres and to Matthias Baas, my gratitude for providing me with technical help on the rendering of the three-dimensional models appearing in the book. To Andrew Beatty at Singular Inversions, appreciation for providing me with a free copy of their programs for the generation of three-dimensional head models. A blossomy ‘whoa’ to Filippo Brunelli, for using these programs to generate the many virtual heads popping up in the figures of the book and feeding some of the algorithms described. To Carla Maria Modena, a lot of thanks for helping in the revision of the manuscript. And, finally, very huge thanks indeed to Tomaso Poggio, the best colleague I ever had, and the main culprit for the appearance of this book, as the first epigraph in the book tells you.

Roberto Brunelli
Trento, Italy

1 INTRODUCTION

> Somewhere, somewhen,
> a two headed strategic meeting
> on face recognition and matters alike:
> t: What about using template matching?
> r: Template matching?
> t: Yes, a simple technique to compare patterns . . .
> r: I'll have a look.
>
> *Faces' faces – r's virtual autobiography*
> ROBERTO BRUNELLI

> Go thither; and, with unattainted eye,
> Compare her face with some that I shall show,
> And I will make thee think thy swan a crow.
>
> *Romeo and Juliet*
> WILLIAM SHAKESPEARE

Computer vision is a wide research field that aims at creating machines that see, not in the limited meaning that they are able to sense the world by optical means, but in the more general meaning that they are able to understand its perceivable structure. Template matching techniques, as now available, have proven to be a very useful tool for this intelligent perception process and have led machines to superhuman performance in tasks such as face recognition. This introductory chapter sets the stage for the rest of the book, where template matching techniques for monochromatic images are discussed.

1.1. Template Matching and Computer Vision

The whole book is dedicated to the problem of template matching in computer vision. While template matching is often considered to be a very basic, limited approach to the most interesting problems of computer vision, it touches upon many old and new techniques in the field.

The two terms *template* and *matching* are used in everyday language, but recalling the definitions more closely related to their technical meaning is useful:

template/pattern

1. Anything fashioned, shaped, or designed to serve as a model from which something is to be made: a model, design, plan, outline.

2. Something formed after a model or prototype, a copy; a likeness, a similitude.
3. An example, an instance; esp. a typical model or a representative instance.

matching

1. Comparing in respect of similarity; to examine the likeness or difference of.

A template may additionally exhibit some variability: not all of its instances are exactly equal (see Figure 1.1). A simple example of template variability is related to its being corrupted by additive noise. Another important example of variability is due to the different viewpoints from which a single object might be observed. Changes in illumination, imaging sensor, or sensor configuration may also cause significant variations. Yet another form of variability derives from intrinsic variability across physical object instances that causes variability of the corresponding image patterns: consider the many variations of faces, all of them sharing a basic structure, but also exhibiting marked differences. Another important source of variability stems from the temporal evolution of a single object, an interesting example being the mouth during speech. Many tasks of our everyday life require that we identify classes of objects in order to take appropriate actions in spite of the significant variations that these objects may exhibit. The purpose of this book is to present a set of techniques by which a computer can perform some of these identifications. The techniques presented share two common features:

- all of them rely on explicit templates, or on representations by which explicit templates can be generated;
- recognition is performed by matching: images, or image regions, are set in comparison to the stored representative templates and are compared in such a way that their appearance (their image representation) plays an explicit and fundamental role.

The simplest template matching technique used in computer vision is illustrated in Figure 1.2. A planar distribution of light intensity values is transformed into a vector $\boldsymbol{x}$ which can be compared, in a coordinate-wise fashion, to a spatially congruent light distribution similarly represented by vector $\boldsymbol{y}$:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 = \frac{1}{N} \, \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{1.1}$$

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{1 + d(\boldsymbol{x}, \boldsymbol{y})}. \tag{1.2}$$

A small value of $d(\boldsymbol{x}, \boldsymbol{y})$ or a high value of $s(\boldsymbol{x}, \boldsymbol{y})$ is indicative of pattern similarity. A simple variation is obtained by substituting the L_2 norm with the L_p norm:

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^p = \frac{1}{N} \, \|\boldsymbol{x} - \boldsymbol{y}\|_p^p. \tag{1.3}$$

If $\boldsymbol{x}$ is representative of our template, we search for other instances of it by superposing it on other images, or portions thereof, searching for the locations of lowest distance $d(\boldsymbol{x}, \boldsymbol{y})$ (or highest similarity $s(\boldsymbol{x}, \boldsymbol{y})$).

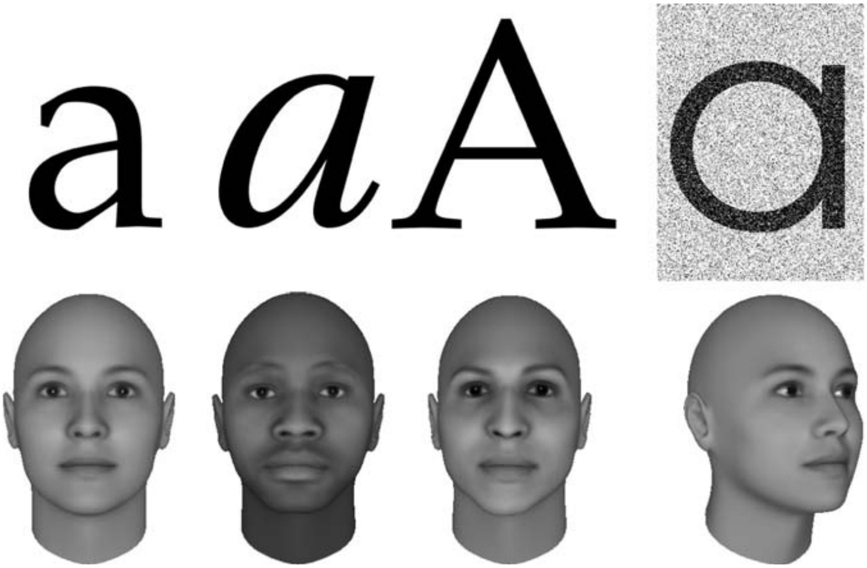


Figure 1.1. Templates from two very common classes: characters and ‘characters’, i.e. faces. Both classes exhibit intrinsic variability and can appear corrupted by noise.

The book shows how this simple template matching technique can be extended to become a flexible and powerful tool supporting the development of sophisticated computer vision systems, such as face recognition systems.

While not a face recognition book, its many examples are related to automated face perception. The main reason for the bias is certainly the background of the author, but there are at least three valid reasons for which face recognition is a valid test bed for template matching techniques. The first one is the widespread interest in the development of high-performing face recognition systems for security applications and for the development of novel services. The second, related reason is that, over the last 20 years, the task has become very popular and it has seen a significant research effort. This has resulted in the development of many algorithms, most of them of the template matching type, providing material for the book. The third reason is that face recognition and facial expression interpretation are two tasks where human performance is considered to be flawless and key to social human behavior. Psychophysical experiments and the evolution of matching techniques have shown that human performance is not flawless and that machines can, sometimes, achieve super human performance.

1.2. The Book

A modern approach to template matching in computer vision touches upon many aspects, from imaging, the very first step in getting the templates, to learning techniques that are key to the possibility of developing new systems with minimal human intervention. The chapters present a balanced description of all necessary concepts and techniques, illustrating them with examples taken from face processing tasks.

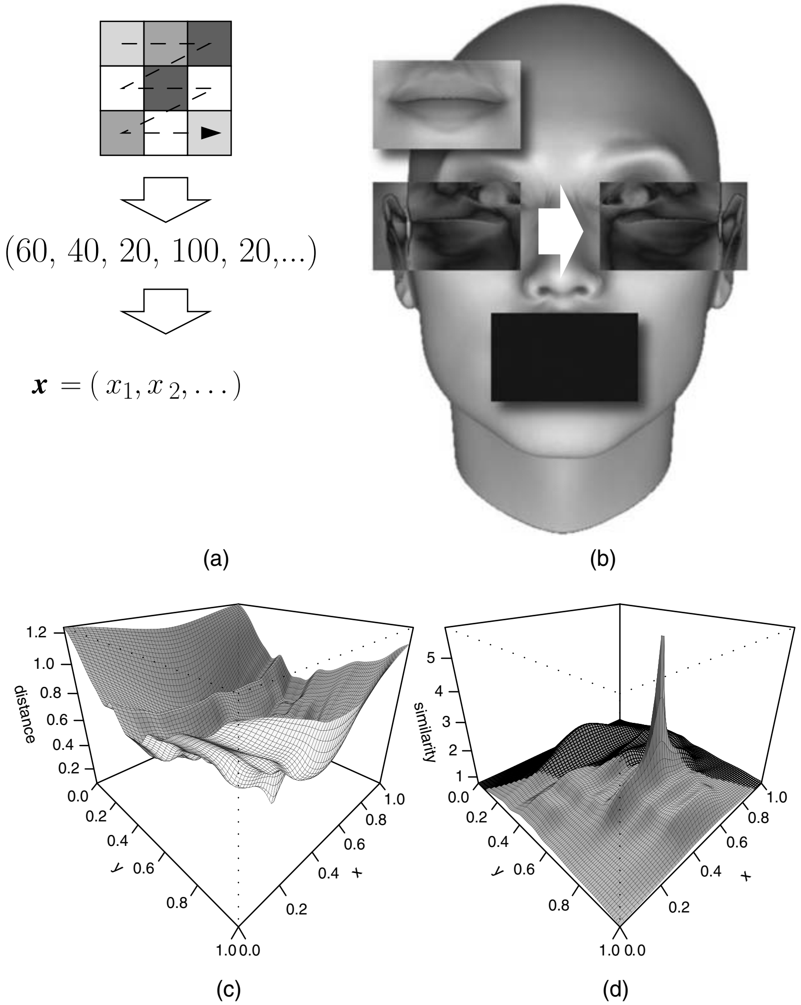


Figure 1.2. The simplest template matching technique: templates are represented as vectors (a) and they are matched by computing their distance in the associated vector space. The template is moved over the image, as a sliding window (b), and the difference between the template and the image is quantified using Equation 1.1, searching for the minimum value (c), or Equation 1.2, searching for the maximum value (d).

A complete description of the imaging process, be it in the case of humans, animals, or computers, would require a (very large) book by itself and we will not attempt it. Chapter 2 discusses some aspects of it that turn out to be critical in the design of artificial vision systems. The basics of how images are created using electromagnetic stimuli and imaging devices are considered. Simple concepts from optics are introduced (including distortion, depth of field, aperture, telecentric lens design) and eyes and digital imaging sensors briefly described. The sampling theorem is presented and its impact on image representation and common image processing operations such as resizing is discussed.

Chapter 3 formally introduces template matching as an hypothesis testing problem. The Bayesian and frequentist approaches are considered with particular emphasis on the Neyman–Pearson paradigm. Matched filters are introduced from a signal processing perspective and simple pattern variability is addressed with the normalized Pearson correlation coefficient. Hypothesis testing often requires the statistical estimation of the parameters characterizing the associated decision function; some subtleties in the estimation of covariance matrices are discussed.

A major issue in template matching is the stability of similarity scores with respect to noise extended to include unmodelled phenomena. Many commonly used estimators suffer from a lack of robustness: small perturbations in the data can drive them towards uninformative values. Chapter 4 addresses the concept of estimator robustness in a technical way, presenting applications of robust statistics to the problem of pattern matching.

Linear correspondence measures like correlation and the sum of squared differences between intensity distributions are fragile. Chapter 5 introduces similarity measures based on the relative ordering of intensity values. These measures have demonstrable robustness both to monotonic image mappings and to the presence of outliers.

While finding a single, well-defined shape is useful, finding instances of a class of shapes can be even more useful. Intraclass variability poses new problems for template matching and several interesting solutions are available. Chapter 6 focuses on the use of projection operators on a one-dimensional space to solve the task. The use of projection operators on multidimensional spaces is covered in Chapter 8.

Finding simple shapes, such as lines and circles, in images may look like a simple task but computational issues coupled with noise and occlusions require some not so naive solutions. In spite of the apparent diversity of lines and areas, it turns out that common approaches to the detection of linear structures can be seen as an efficient implementation of matched filters. Chapter 7 describes how to compute salient image discontinuities and how simple shapes embedded in the resulting map can be located with the Radon/Hough transform.

The representation of images of even moderate resolution requires a significant amount of numeric data, usually one (three) values per pixel if the typical array-based method is adopted. Chapter 8 investigates the possibility of alternative ways of representing iconic data so that a large variety of images can be represented using vectors of reduced dimensionality. Besides significant storage savings, these approaches provide significant benefits to template detection and recognition algorithms, improving their efficiency and effectiveness.

Chapter 9 addresses a couple of cases that are not easily reduced to pattern detection and classification. One such case is the detailed estimation of the parameters of a parametric curve: while Hough/Radon techniques may be sufficient, accurate estimation may benefit from specific approaches. Another important case is the comparison of anatomical structures, such as brain sections, across different individuals or, for the same person, over time. Instead of modeling the variability of the patterns within a class as a static multidimensional manifold, we may focus on the constrained deformation of a parameterized model and measure similarity by the deformation stress.

The drawback of template matching is its high computational cost which has two distinct origins. The first source of complexity is the necessity of using multiple templates to accommodate the variability exhibited by the appearance of complex objects. The second source of complexity is related to the representation of the templates: the higher the resolution, i.e. the number of pixels, the heavier the computational requirements. Besides some computational tricks, Chapter 10 presents more organized, structural ways to improve the speed at which template matching can be performed.

Matching sets of points using techniques targeted at area matching is far from optimal, with regard to both efficiency and effectiveness. Chapter 11 shows how to compare sparse templates, composed by points with no textural properties, using an appropriate distance. Robustness to noise and template deformation as well as computational efficiency are analyzed.

When the probability distribution of the templates is unknown, the design of a classifier becomes more complex and many critical estimation issues surface. Chapter 12 presents basic results upon which two interrelated, powerful classifier design paradigms stand: regularization networks and support vector machines.

Many applications in image processing rely on robust detection of image features and accurate estimation of their parameters. Features may be too numerous to justify the process of deriving a new detector for each one. Chapter 13 exploits the results presented in Chapter 8 to build a single, flexible, and efficient detection mechanism. The complementary aspect of detecting templates considered as a set of separate features will also be addressed and an efficient architecture presented.

Template matching techniques are a key ingredient of many computer vision systems, ranging from quality control to object recognition systems among which biometric identification systems have today a prominent position. Among biometric systems, those based on face recognition have been the subject of extensive research. This popularity is due to many factors, from the non-invasiveness of the technique, to the high expectations due to the widely held belief that human face recognition mechanisms perform flawlessly. Building a face recognition system from the ground up is a complex task and Chapter 14 addresses all the required practical steps: preprocessing issues, feature scoring, the integration of multiple features and modalities, and the final classification stage.

The process of developing a computer vision system for a specific task often requires the interactive exploration of several alternative approaches and variants, preliminary parameter tuning, and more. Appendix A introduces AnImAl, an image processing package written for the R statistical software system. AnImAl, which relies on an algebraic formalization of the concept of image, supports interactive image processing by adding to images a self-documenting capability based on a history mechanism. The documentation facilities of the resulting interactive environment support a practical approach to reproducible research.

A key need in the development of algorithms in computer vision (as in many other fields) is the availability of large datasets for training and testing them. Ideally, datasets should cover the expected variability range of data and be supported by high-quality annotations describing what they represent so that the response of an algorithm can be compared to reality. Gathering large, high-quality datasets is, however, time consuming. An alternative is available for computer vision research: computer graphics systems can be used to generate photorealistic images of complex environments together with supporting ground truth information. Appendix B shows how these systems can be exploited to generate a flexible (and cheap) evaluation environment.

Evaluation of algorithms and systems is a complex task. Appendix C addresses four related questions that are important from a practical and methodological point of view: what is a good response of a template matching system, how can we exploit data to train and at the same time evaluate a classification system, how can we describe in a compact but informative way the performance of a classification system, and, finally, how can we compare multiple classification systems for the same task in order to assess the state of the art of a technology?

The exposition of the main chapter topics is complemented by several intermezzos which provide ancillary material or refresh the memory of useful results. The arguments presented are illustrated with several examples from a very specific computer vision research topic: face detection, recognition, and analysis. There are three main reasons for the very biased choice: the research background of the author, the relevance of the task in the development of biometrics systems, and the possibility that a computational solution to these problems helps understanding (and benefits from the understanding of) the way people do it. Some of the images appearing in the book are generated using the computer graphics techniques described in Appendix B and the packages POV-ray (The Povray Team 2008), POVMan (Krouverk 2005), Aqsis (The Aqsis Team 2007), and FaceGen (Singular Inversions 2008).

Intermezzo 1.1. The definition of intermezzo

intermezzo

1. A brief entertainment between two acts of a play; an entr'acte.
2. A short movement separating the major sections of a lengthy composition or work.

References to relevant literature are not inserted throughout chapter text but are postponed to a final chapter section. Their order of presentation follows the structure of the chapter. All papers on which the chapter is based are listed and pointers to additional material are also provided. References are not meant to be exhaustive, but the interested reader can find additional literature coverage in the cited papers.

1.3. Bibliographical Remarks

This book, while addressing a very specific technique of computer vision, touches upon concepts and methods typical of other fields, from optics to machine learning and its comprehension benefits from readings in these fields.

Among the many books on computer vision, the one by Marr (1982) is perhaps the most fascinating. The book by Horn (1986) still provides an excellent introduction to the fundamental aspects of computer vision, with a careful treatment of image formation. A more recent book is that by Forsyth and Ponce (2002).

Basic notions of probability and statistics can be found in Papoulis (1965). Pattern classification is considered in detail in the books by Fukunaga (1990) and Duda *et al.* (2000). Other important reference books are Moon and Stirling (2000) and Bishop (2007).

A very good, albeit concise, reference for basic mathematical concepts and results is the *Encyclopedic Dictionary of Mathematics* (Mathematical Society of Japan 1993). A wide coverage of numerical techniques is provided by Press *et al.* (2007).

Two interesting papers on computer and human face recognition are those by Sinha *et al.* (2006) and O'Toole *et al.* (2007). The former presents several results on human face analysis processes that may provide guidance for the development of computer vision algorithms. The latter presents some results showing that, at least in some situations, computer vision efforts resulted in algorithms capable of superhuman performance.

References

Bishop C 2007 *Pattern Recognition and Machine Learning*. Springer.

Duda R, Hart P and Stork D 2000 *Pattern Classification* 2nd edn. John Wiley & Sons, Ltd.

Forsyth D and Ponce J 2002 *Computer Vision: A Modern Approach*. Prentice Hall.

Fukunaga K 1990 *Statistical Pattern Recognition* 2nd edn. Academic Press.

Horn B 1986 *Robot Vision*. MIT Press.

Krouverk V 2005 POVMan v1.2. http://www.aetec.ee/fv/vkhomep.nsf/pages/povman2.

Marr D 1982 *Vision*. W.H. Freeman.

Mathematical Society of Japan 1993 *Encyclopedic Dictionary of Mathematics* 2 edn. MIT Press.

Moon T and Stirling W 2000 *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall.

O'Toole A, Phillips P, Fang J, Ayyad J, Penard N and Abdi H 2007 Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(9), 1642–1646.

Papoulis A 1965 *Probability, Random Variables and Stochastic Processes*. McGraw-Hill.

Press W, Teukolsky S, Vetterling W and Flannery B 2007 *Numerical Recipes* 3rd edn. Cambridge University Press.

Singular Inversions 2008 FaceGen Modeller 3.2. http://www.facegen.com.

Sinha P, Balas B, Ostrovsky Y and Russell R 2006 Face recognition by humans: nineteen results all computer vision researchers should know about. *Proceedings of the IEEE* **94**, 1948–1962. Face Recognition.

The Aqsis Team 2007 Aqsis v1.2. http://www.aqsis.org/.

The Povray Team 2008 The Persistence of Vision Raytracer v3.6. http://www.povray.org/.

2 THE IMAGING PROCESS

> I have a good eye, uncle; I can see a church by daylight.
>
> *Much Ado About Nothing*
> WILLIAM SHAKESPEARE

A complete description of the imaging process, be it in the case of humans, animals, or computers, would require a (very large) book by itself and we will not attempt it. Rather, we discuss some aspects of it that turn out to be critical in the design of artificial vision systems. The basics of how images are created using electromagnetic stimuli and imaging devices will be considered. Simple concepts from optics will be introduced (including distortion, depth of field, aperture, telecentric lens design). Eyes and digital imaging sensors will be briefly considered. The sampling theorem is presented and its impact on image representation and common image processing operations such as resizing is discussed.

2.1. Image Creation

Computer vision can be considered as the science and technology of machines that see, obtaining information on the real world from images of it. Images are created by the interaction of light with objects and they are captured by optical devices whose nature may differ significantly.

2.1.1. LIGHT

What is commonly understood by light is actually a propagating oscillatory disturbance in the electromagnetic field which describes the interaction of charged particles, such as electrons. The field derives from the close interaction of varying electric and magnetic fields whose coupling is described by a set of partial differential equations known as Maxwell's equations. An important consequence of them is the second-order partial differential equation that describes the propagation of electromagnetic waves through a medium or in vacuum. The free space version of the electromagnetic wave equation is

$$\left(\nabla^2 - \frac{1}{c}\frac{\partial^2}{\partial t^2}\right)\boldsymbol{E} = 0 \tag{2.1}$$

where $\boldsymbol{E}$ is the electric field (and similarly for the magnetic field $\boldsymbol{B}$): light is a visible solution of this equation. From the theory of Fourier decomposition, the finite spatio-temporal extent of real physical waves results in the fact that they can be described as a superposition of an infinite set of sinusoidal frequencies. In many cases we may then limit our analysis to pure

Template Matching Techniques in Computer Vision: Theory and Practice Roberto Brunelli

Intermezzo 2.1. Convolution and its properties

As convolution, and the closely related cross-correlation, play a significant role in our analysis of template matching techniques, it is useful to recall their definitions and their most important properties. Both operations feature a continuous and a discrete definition. The convolution of two real continuous functions f and g is a new function defined as

$$(f * g)(x) = \int f(y)g(x-y)\,dy \tag{2.2}$$

where integration is extended to the domain over which the two functions are defined. The cross-correlation is a new function defined as

$$(f \otimes g)(x) = \int f(y)g(x+y)\,dy. \tag{2.3}$$

The discrete versions are

$$(f * g)(i) = \sum_j f(j)g(i-j) \tag{2.4}$$

$$(f \otimes g)(i) = \sum_j f(j)g(i+j). \tag{2.5}$$

The two operations provide the same result when one of the two argument functions is even. The main properties of convolution are

$$f * g = g * f \tag{2.6}$$

$$f * (g + h) = (f * g) + (f * h) \tag{2.7}$$

$$f * (g * h) = (f * g) * h \tag{2.8}$$

$$(f * g)(x - a) = f(x - a) * g(x) \tag{2.9}$$

$$\frac{d}{dx}(f * g)(x) = \left(\frac{d}{dx}f\right) * g \tag{2.10}$$

$$\text{support}(f * g) = \text{support}(f) \cup \text{support}(g). \tag{2.11}$$

Convolution and correlation can be extended in a straightforward way to the multidimensional case.

sinusoidal components that we can write conveniently in complex form, remembering that we should finally get the real or imaginary part of the complex results:

$$\boldsymbol{E}(\boldsymbol{x}, t) = \boldsymbol{E}_0 e^{\boldsymbol{k}\cdot\boldsymbol{x} - \omega t} \tag{2.12}$$

where $\boldsymbol{k}$ is the wave vector representing the propagating direction and the angular frequency ω is related to frequency f by $\omega = 2\pi f$. The electric and magnetic fields for the plane wave represented by Equation 2.12 are perpendicular to each other and to the direction of propagation of the wave. The velocity of the wave c, the wavelength λ, the angular frequency ω, and the size of the wave vector $\boldsymbol{k}$ are related by

$$c = \frac{\omega}{k} = \frac{\omega\lambda}{2\pi} = f\lambda. \tag{2.13}$$

Besides wavelength, two other properties of light are of practical importance: its polarization and its intensity. Polarization and its usefulness in imaging are considered in Intermezzo 2.2. Even if we described light using the wave equation, there are no sensors to detect directly its amplitude and phase. The only quantity that can be detected is the intensity I of the radiation incident on the sensor, the irradiance; that is, the time average of the radiation energy which

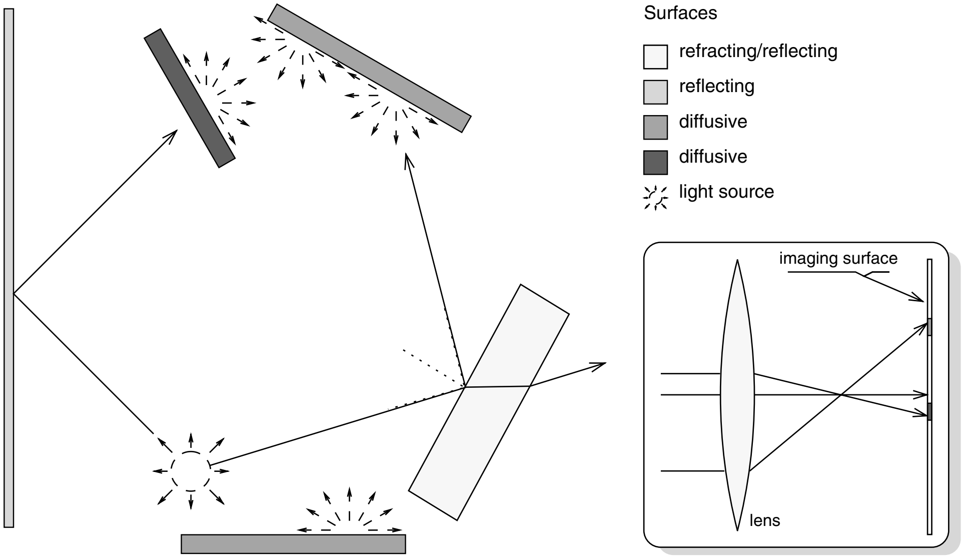


Figure 2.1. The process of light propagation. The goal of computer vision is to deduce the path of light from its observation at a sensing surface.

crosses unit area in unit time. For a plane wave

$$I(x_1, x_2) \propto E_0^2 \tag{2.14}$$

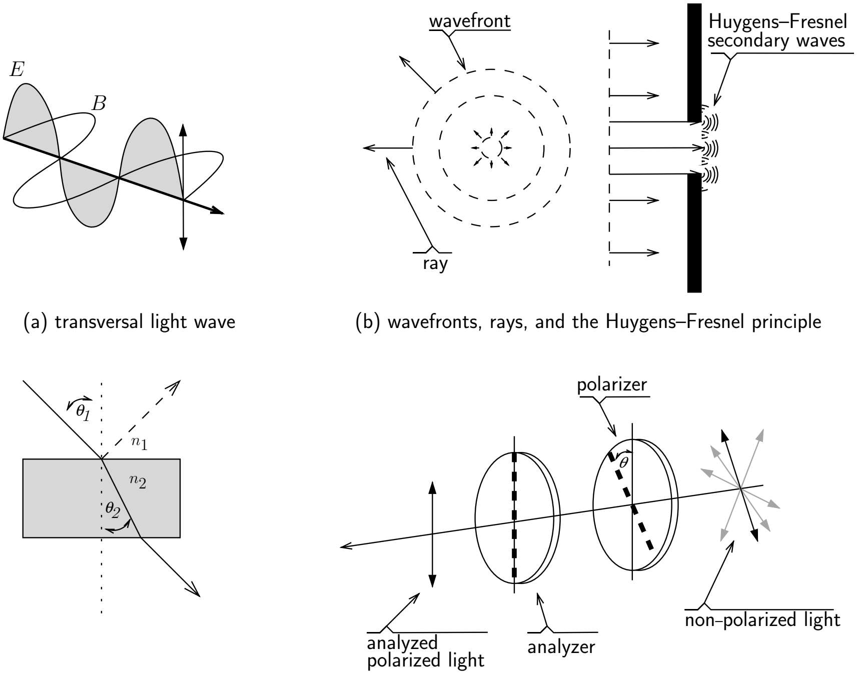
where (x_1, x_2) represents a point on the (real or virtual) surface at which we perform the measurement (see Figure 2.1). Deriving information on the world from the two-dimensional intensity map $I(x_1, x_2)$ is the goal of computer vision. Plane waves are not the only solution to Maxwell's equations, spherical waves being another very important one. A spherical wave is characterized by the fact that its components depend only on time and on the distance r from its center, where the light source is located (see Intermezzo 2.3). Energy conservation requires the irradiance of a spherical wave to decay as r^{-2}, a fact often appreciated when photographing with a flash unit. A spherical wave can be approximated with a plane wave when r is large; in many cases of interest plane waves can then provide a good approximation of light waves. Two entities (see Figure 2.2) are fundamental for studying light propagation:

Definition 2.1.1. *A wavefront is a surface over which an optical disturbance has a constant phase.*

Definition 2.1.2. *Rays are lines normal to the wavefronts at every point of intersection.*

The discovery of the photoelectric effect, by which light striking a metal surface ejects electrons whose energy is proportional to the frequency and not the intensity of light, led to quantum field theory: interactions among particles are mediated by other particles, the photon being the mediating particle for the electromagnetic field. Photons have an associated energy

$$E_\lambda = \frac{hc}{\lambda} \tag{2.15}$$



(a) transversal light wave

(b) wavefronts, rays, and the Huygens–Fresnel principle

(c) Snell's law

(d) analysis of polarized light

Figure 2.2. A graphical illustration of some important optics concepts described in the chapter.

where h is Planck's constant: from a particle point of view, intensity is related to the number of photons. As we will see, both wave and particle aspects of light have important consequences in the development of systems that sense the world using electromagnetic radiation, and they fix some fundamental limits for them. The speed of light depends on the medium of propagation, and in a linear, isotropic, and non-dispersive material is

$$c = \frac{c_0}{n} \tag{2.16}$$

where n is the refractive index of the medium and c_0 is the speed of light in vacuum. Usually $n > 1$ and it depends on frequency: it generally decreases with decreasing frequency (increasing wavelength). When light crosses the boundary between media with different refractive indexes it changes its direction and is partially reflected. These effects allow us to control the propagation of light by interposing properly shaped elements of different refractive indexes. The refraction of light crossing the boundary of two different isotropic media results in a change of the propagation direction obeying Snell's law

$$n_1 \sin\,\theta_1 = n_2 \sin\theta_2 \tag{2.17}$$

where θ_1 represents the angle with respect to the normal of the boundaries: when $n_2 > n_1$ the light will be deflected towards the normal (see Figure 2.2). When $n_1 > n_2$, so that light passes

Intermezzo 2.2. Polarized light

The vector representing the electric field can be decomposed into two orthogonal components. In the case of a simple harmonic wave the two components have the same frequency but may have different phases. However, if they have the same phase, the direction of the electric vector remains constant and its changes are restricted to a constant plane: the light is linearly polarized. Polarized light may result from the reflection of unpolarized light from dielectric materials (as electric dipoles do not emit in the direction along which they oscillate) or from selective transmission of one of the two components of the electric field (e.g. using a Glan–Thomson prism). A natural cause of partially polarized light is light scattering by small particles. Polarized light is important in computer vision because it can be selectively filtered using polarizers according to Malu's law:

$$I = I_0 \cos^2 \theta_i \tag{2.18}$$

where I_0 is the intensity of the polarized light, I the intensity transmitted by the filter, and θ_i is the angle of the polarized light to that of the polarizer (see Figure 2.2). Reflection preserves polarization while diffusion by scattering surfaces does not, producing instead unpolarized light. Let us consider an inspection problem where the integrity of a non-metallic marker overlying a metallic object must be verified. The visible metallic parts of the object produce a lot of glare that may severely impair the imaging of the marker. However, if we illuminate the specimen with polarized light and we observe the scene with a properly rotated polarizer we may get rid of the reflections from the metal parts (suppressed by Malu's law) while still perceiving a fraction of the unpolarized light from the diffusing surfaces. Another case is the inspection of specimens immersed in water. In this case the water surface reflects partially polarized light. Polarization of reflected light is complete at the Brewster angle $\theta_B = \arctan(n_2/n_1)$, n_2, n_1 being the refractive indexes of the materials, that for visible light is approximately 53° to the normal for an air–water interface. These reflections, which would prevent a clear image of objects below the surface, can be reduced using a properly oriented polarizer.

The detection of reflected light polarization is of help in several image understanding applications, including the discrimination of dielectric/metal material, segmentation of specularities, and separation of specular and diffuse reflection components (Wolff 1995).

from a dense to a less dense medium, e.g. from water to air, we may observe total internal reflection: no refracted ray exists.

Different frequencies of oscillation give rise to the different forms of electromagnetic radiation, from radio waves at the lowest frequencies, to visible light at the intermediate frequencies, to gamma rays at the highest frequencies. The whole set of possibilities is known as the electromagnetic spectrum and the nomenclature for the main portions is reported in Table 2.1. The study of the properties and behavior of visible light, with the addition of infrared and ultraviolet light, and of its interaction with matter is the subject of optics, itself an important area of physics. As light is an electromagnetic wave, similar phenomena occur over the complete electromagnetic spectrum and can also be found in the analysis of elementary particles due to wave–particle duality, the fact that matter exhibits both wave-like and particle-like properties.

2.1.2. GATHERING LIGHT

Images are created by controlling light propagation with optical devices so that we can effectively detect its intensity without disrupting the information it contains on the world through which it traveled. Refraction of light by means of media with different refractive indexes, the lenses, is the basis of optical systems. Interposition of glass elements of different shapes and refractive indexes, allows us to control the propagation of light so that different rays emitted by a single point in the world can be focused into a corresponding image of

Table 2.1. The different portions of the electromagnetic spectrum. The most common unit for wavelength is the angstrom (10^{-10} m).

Region	Frequency (Hz)	Wavelength (Å)
Radio	$< 3 \times 10^9$	$> 10^9$
Microwave	3×10^9 to 3×10^{12}	10^9–10^6
Infrared	3×10^{12} to 4.3×10^{14}	10^6–7000
Visible	4.3×10^{14} to 7.5×10^{14}	7000–4000
Ultraviolet	7.5×10^{14} to 3×10^{17}	4000–10
X-rays	3×10^{17} to 3×10^{19}	10–0.1
Gamma rays	$> 3 \times 10^{19}$	< 0.1

Intermezzo 2.3. Light sources

There are many sources of light that we encounter daily, such as the sun, incandescent bulbs, and fluorescent tubes. The basic mechanism underlying light emission is atomic excitation and relaxation. When an atom adsorbs energy, its outer electrons move to an excited state from which they relax to the ground state, returning the energy in the form of photons, whose wavelength is related to their energy. The specification of the spectral exitance or spectral flux density, the emitted power per unit area per unit wavelength interval, characterizes a light source. The human eye is sensitive to a limited range of electromagnetic radiation and it perceives different wavelengths as different colors: the longer wavelengths as red, the shorter ones as blue.

Excitation by thermal energy results in the emission of photons of all energies, and the corresponding light spectrum is continuous. The finite time over which all electron transitions happen and atomic thermal motion are responsible for the fact that light emission is never perfectly monochromatic but characterized by a frequency bandwidth $\Delta\nu$ inversely proportional to the temporal extent of the events associated with the emission

$$\Delta\nu \approx \frac{1}{\Delta\tau_c}$$

where $\Delta\tau_c$ is also known as coherence time. For a blackbody, i.e. a perfect absorber, Planck's radiation law holds

$$I(\lambda, t) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc(\lambda kt)^{-1}} - 1}$$

and the wavelength of maximum exitance λ_{peak} obeys Wien's law

$$\lambda_{\text{peak}}\, T = 2.9 \times 10^6 \text{ nm K}$$

so that the hotter the body, the bluer the light. Thermal light sources can then be indexed effectively by their temperature.

When light passes through an absorbing medium, each wavelength may undergo selective absorption: the spectral distribution of light changes. The same thing happens by selective reflection from surfaces; this is the reason why we see them in different colors. In most cases, light-detecting devices integrate light of several wavelengths using a convolution mechanism: they compute a spectrally weighted average of the incoming light. Multiple convolution kernels may operate on different spectral regions, resulting in multichannel, or color, imaging. Different sensors are characterized by different kernels so that their colors may differ: images from different sensors need to be calibrated so that they agree on the color they report. When a single kernel is used, we have monochromatic vision. The light spectrum that a sensor perceives depends then on the light sources and on the media and objects the light interacts with. It is sometimes possible to facilitate the task of computer vision by means of appropriate lighting and environment colors.

the point on a light-detecting device. A typical configuration is composed of several radially symmetric lenses whose symmetry axes are all aligned along the optical axis of the system.

Computations are quite complex in the general case but may be simplified significantly by considering the limit of small angles from the optical axis (the paraxial approximation) and, at the same time, thin lenses with spherical surfaces immersed in air (see Figure 2.3). The main property of a lens is its focal length f, the distance on the optical axis from its center to a point onto which collimated light parallel to the axis is focused:

$$\frac{1}{f} \approx (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{2.19}$$

where R_1 (R_2) is the radius of curvature of the lens surface closest to (farthest from) the light source, and n is shorthand for $n_{lm} = n_l/n_m$, the relative refraction index of the lens (n_l) and the medium where the lens is immersed (n_m). The formula can be derived using the approximation $\sin\theta \approx \theta$ in Snell's law and considering spherical lenses. The focal length and the dimensions $w \times h$ of the imaging surface determine the field of view (FOV) of the optical system:

$$f = \frac{w}{h} \Big/ \tan\left(\frac{\text{FOV}}{2}\right). \tag{2.20}$$

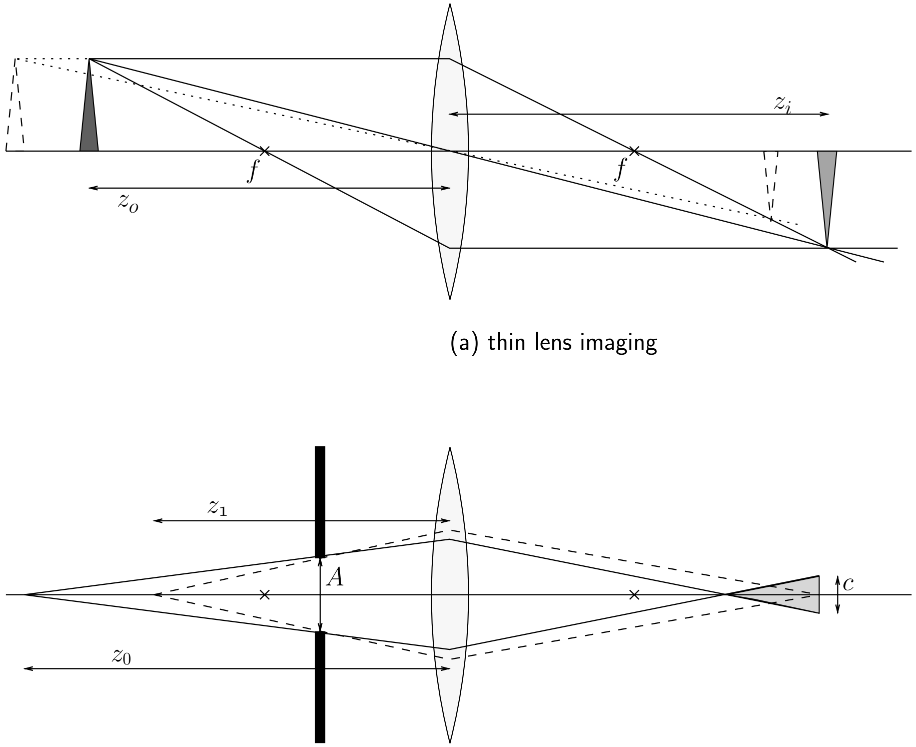
The reciprocal of the focal length is its optical power. Under the approximations considered the thin lens equation relates the distance z_o of an object from the lens to the distance z_i at which the object is focused

$$\frac{1}{z_o} + \frac{1}{z_i} = \frac{1}{f}. \tag{2.21}$$

This can be most easily seen by considering two rays emanating from points on the object: a ray parallel to the optical axis, which when refracted passes through the focal point, and one passing undeflected through the optical center of the lens (the chief ray). If we put a screen at distance z_i from the lens, a point source at z_o will be imaged as a point on the screen, but if the screen is moved closer (or farther) from the lens, the image of the point is transformed into a disc. The diameter c of the circle corresponding to the out of focus imaging of a point source can be determined using the thin lens equation (see also Figure 2.3) and turns out to be

$$c = A \cdot \frac{|z_o - z_1|}{z_o} \cdot \frac{f}{z_1 - f} = \frac{|z_o - z_1|}{z_o} \cdot \frac{f^2}{N(z_1 - f)} \tag{2.22}$$

where z_o is the distance of the object from the lens, z_1 is the distance for which the lens is focused, A is the diameter of the incident light beam as it reaches the lens, the so-called lens aperture, and $N = f/A$ is the so-called f-number commonly used in photography. The effective f-number N_e is defined as the ratio of the height to the diameter of the light cone that emerges from the lens. The f-number is a measure of the light-gathering capability of a lens. A way to reduce the diameter of the circle of confusion is then to reduce the aperture of the optical system. As we will see, due to diffraction effects, there is a limiting useful aperture size below which image quality will deteriorate (see Section 2.1.3). A quantity related to the circle of confusion is the depth of field, a measure of the range over which the circle of confusion is below some critical value. Another drawback of reducing the aperture of the optical system is that the amount of light gathered by the system will decrease, resulting in longer exposure time and increased noise (see Section 2.1.4 and Section 2.3). The pinhole



(a) thin lens imaging

(b) circle of confusion geometry

Figure 2.3. A basic optical system: the working of a lens under paraxial approximation (a) and the geometry of the circle of confusion for out of focus imaging (b).

camera, a lens-less optical device described in Figure 2.8 and Intermezzo 2.4, is the simplest optical system and it is often used to model the imaging process in computer graphics and computer vision.

The image produced by a point source is called the point spread function (PSF), and we can represent it as a function of the coordinates on the image plane $\mathrm{PSF}(x, y)$. If the only effect of translating a point in the object plane is a proportional translation of the PSF in the image plane, the system is said to be isoplanatic. When image light is incoherent, without a fixed phase relationship, optical systems are linear in intensity. An important consequence is that the image obtained from a linear isoplanatic system is the convolution (see Intermezzo 2.1) of its PSF with the object plane image:

$$I_{\text{image}}(x, y) = \iint I_{\text{object}}(x', y')\mathrm{PSF}(x - x', y - y')\, dx'\, dy' = (I_{\text{object}} * \mathrm{PSF})(x, y). \tag{2.23}$$

If we consider an off-axis point source P with θ the angle between the optical axis and the principal ray from the point through the center of the aperture, the rays emanating from P will see a foreshortened aperture due to the angular displacement. As a consequence, the light-gathering area is reduced for them, resulting in a falling off of the irradiance I on the sensor with respect to the radiance L on the surface in the direction of the lens:

$$I = L \frac{\pi}{4} \frac{\cos^4 \theta}{N_e^2}. \tag{2.24}$$

This effect is responsible for the radial intensity falloff (vignetting) exhibited by wide-angle lenses.

The simple optical systems considered so far produce perspective images. The image dimensions of equally sized objects are inversely proportional to the corresponding object distances (see Figure 2.3). This variability adversely affects the capability of a computer vision system to recognize objects as it must account for their varying size: the appearance of an object depends on its position within the imaged field and on its distance from the lens. Coupling of simple optical systems and judicious insertion of apertures allow us to build a telecentric system that can produce orthographic images: the appearance of the object does not depend on its distance from the lens or on its position within the field of view (see Figure 2.4). The telecentric design exploits the fact that rays passing through the focal point must emerge (enter) parallel to the optical axis. Let us consider the optical system obtained by removing lens L_2 substituting it with the image screen. The aperture stop limits the bundle of rays and the central ray, the one passing through the focal point, is parallel to the optical axis (on the object side). If we move the object towards the camera, the central ray will remain the same and it will reach the screen at the same position: the size of the object does not change. However, if we move the object it goes out of focus as can be easily seen by considering a point on the optical axis. In a telecentric system, in fact, the aperture, besides controlling the ray bundle to obtain an orthographic projection, continues to control the circle of confusion. If we want to focus the new position we must move the screen, but this would change the size of the object. The introduction of the second lens makes the size invariant also to the changes in screen position necessary to focus objects at different distances as the central ray exits from L_2 parallel to the optical axis. There is an additional, important advantage in using an image side telecentric design: the irradiance of the sensor plane does not change when we change the focusing distance. The reason is that the angular size of the aperture seen by a point on the image plane is constant due to the fact that rays passing through the same point on the focal plane (where the stop is placed) emerge parallel to each other on the image side of the lens. As a consequence, the effective f-number does not change and the intensity on the image plane according to Equation 2.24 does not change. A major limitation of telecentric lenses is that they must be as large as the largest object that needs to be imaged. The effects of focal length and projection type are illustrated in Figure 2.5.

The treatment so far was based on the paraxial approximation that in many real-life situations is not valid. In these cases, an optical system designed using the paraxial approximation exhibits several distortions whose correction requires an approximation up to third order for the $\sin \theta$ function appearing in Snell's law

$$\sin \theta \approx \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!}. \tag{2.25}$$

Distortion of first-order optics can be divided into two main groups: monochromatic (Seidel) aberrations and chromatic aberrations, the latter related to the fact that lenses bring different colors of light to a focus at different points as the refractive index depends on wavelength.

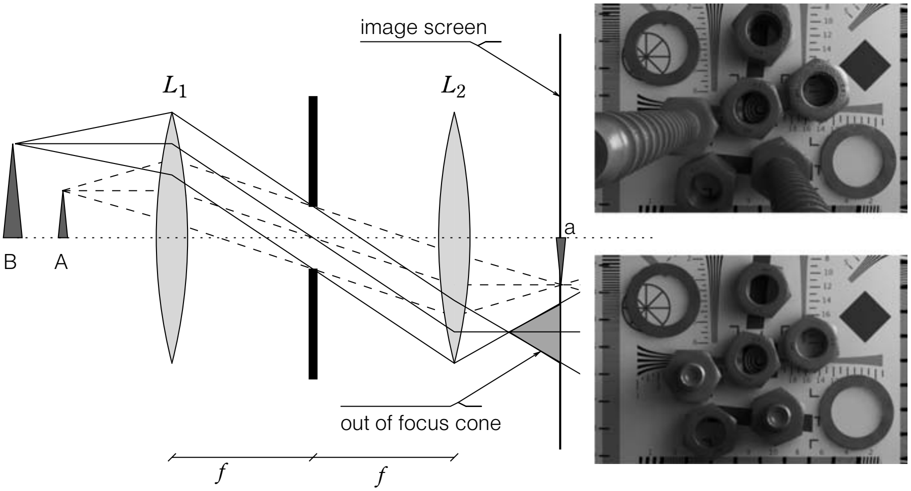


Figure 2.4. A doubly telecentric system obtained by using two lenses with one focal point in common and an aperture positioned at it. The two (synthetic) images on the right show the difference between perspective imaging (top) and telecentric imaging (bottom).

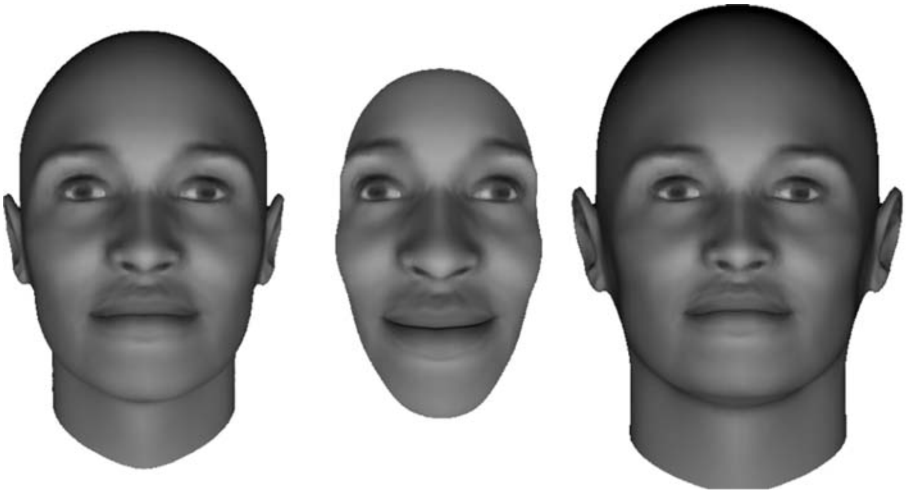


Figure 2.5. Camera focal length and projection type have a significant impact on the appearance of objects. From left to right: perspective projections with a field of view of 20° and 60° respectively, and orthographic projection.

Among the monochromatic aberrations we consider only those related to geometrical distortions, the most important in the field of computer vision. The most common cause for distortion is the introduction of a stop, often needed to correct other aberrations (see Figure 2.6). The position of the stop influences the path of the chief ray, the ray passing through the center of the stop. If the stop is positioned at the lens, it will not be refracted and it will leave the lens without changing its angle: the system is orthoscopic and exhibits no