An Invitation to Image Analysis and Pattern Recognition

Fred A. Hamprecht
University of Heidelberg

Version of September 26, 2010

There are hence many unresolved forward references to yet unwritten parts (there be dragons). I make part 1 available anyway in the hope that it may be useful.

1 Introduction

1.1 Overview

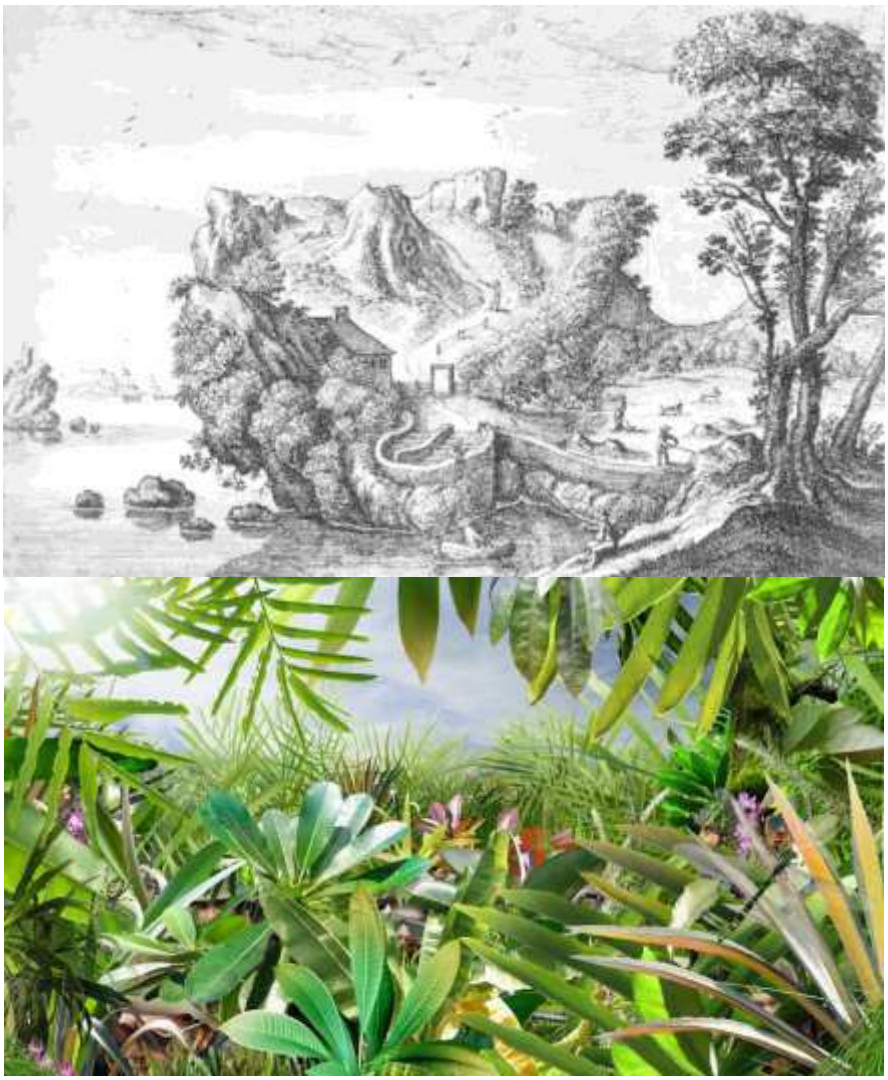
This script will take you on a guided tour to the intriguing world of "pattern recognition": the art and science of analyzing data. Our focus will be the automated analysis of objects or processes whose information content and complexity at least partly reside in their spatial layout. But before delving into technical details, this first chapter will take a step back, allowing to put things into perspective: to remember what the value of vision is (section 1.2), to argue why computer vision is an important and a difficult subject (section 1.3), and to summarize the categories of tasks that will be treated in this script (section 1.4).

1.2 Animate vision

We tend to not think about it, but: vision is almost too good to be true! It is indeed a magnificent sense which transcends the narrow confines of our own physical extent; which is, by all means and purposes, instantaneous; and which allows us, "in the blink of an eye", to make sense of complex scenes and react adequately. The reason we tend not to think about it is that we are endowed with vision almost from the moment we are born; and that it is an unconscious activity: we cannot help but see.

Actually, we compose, rather than perceive, a visual picture of our world: one token of evidence is given by geometrical optics which shows that the world is projected on our retina upside down. This is where our light receptors lie and yet we have the impression that the world is standing "on its feet". This seeming contradiction was cited to refute Johannes Kepler's proposal that the eye acts like a camera obscura. More evidence for the active construction of our percepts is given by optical illusions: these provide a vivid demonstration of the fact that our innate or acquired vision mechanisms help explain the world most of the time, but do sometimes lead us astray, see Figs. 1.1, 1.3.

Hence vision is an "active", if involuntary, achievement in the perceptual sense, but "passive" in a physical sense: we simply capitalize on the copious number of photons reflected or emitted by an object. The latter notion is relatively recent: from the times of Greek physician Empedocles until the 17th century or so, the prevailing belief was that the eye is endowed with



1 Introduction

Figure 1.1: Sample of Renaissance metamorphic mannerism, by Wenceslaus Hollar (1607-1677). It shows that what we see is not what we get: sometimes more, as in this case, and sometimes less, see next figure.

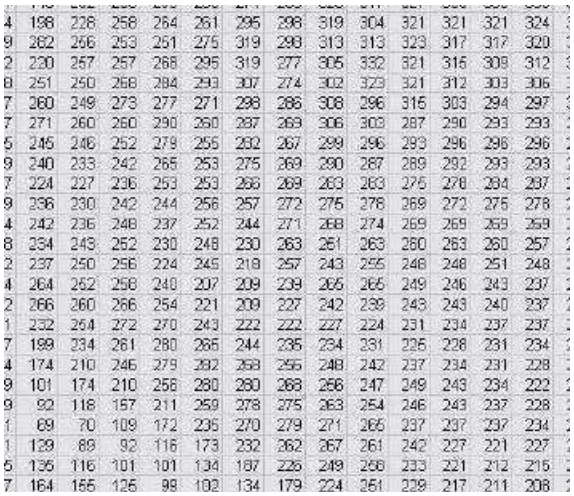
Figure 1.2: Camouflage picture of at least eight commandos by contemporary artist Derek Bacon.

Figure 1.3: Example of false perspective masterpiece created by Francesco Borromini (1599-1667). A rising floor and other cues insinuate a life sized statue at the end of the gallery. In actual fact, the statue rises to a mere 60cm. an “internal fire” and also emits some kind of radiation, rather than merely records incident light [11].

Human vision offers great spatial resolution; and while some species have evolved alternate senses (think bats), vision is unparalleled in the velocity of its stimuli – the speed of light – and in its range. It is only too natural, then, that one would try and endow computers or machines with at least a rudimentary form of vision. However, the true complexity of this exercise that we perform continuously and ostensibly without effort becomes apparent when trying to emulate some form of visual perception in computers: it turns out to be a formidable task, as anybody who has ever tried will confirm.

1.3 Why image processing is difficult

To us, having a picture amounts to understanding it. Admittedly, interpretations often differ between individuals, and surely you have experienced examples where your own perception of an image has changed over time, as a function of additional knowledge or experiences acquired. This perceived immediacy of our own image understanding often leads us to underestimate

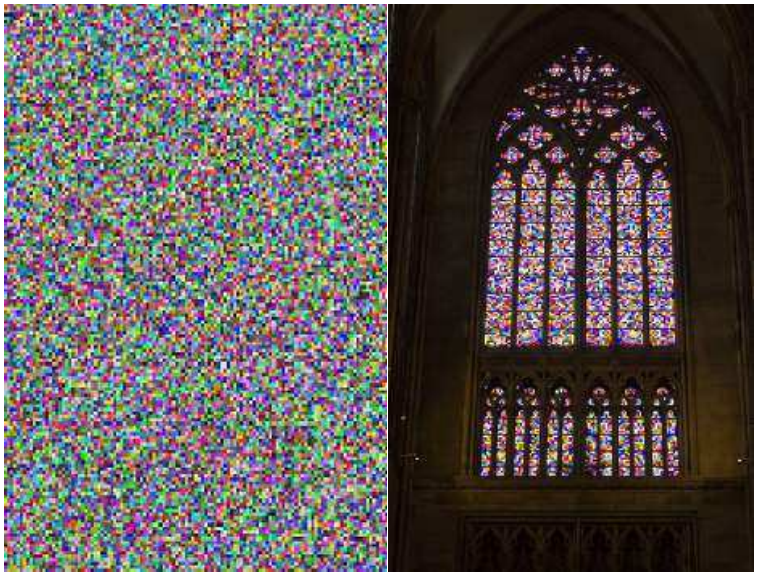


1 Introduction

Figure 1.4: This is how a computer "perceives" an image: as a bunch of numbers. The complexity of image analysis, let alone scene understanding. In contrast to ourselves, the computer does not "understand" a digital picture that it stores with so much more perfection than we ever can. To the computer, any picture looks as shown in Fig. 1.4. As far as a computer is concerned, some images may allow for a higher ratio of compression than others, but other than that they are all equally meaningless.

Assume, for a moment, that our task is to write an algorithm that counts the number of cells labeled with a fluorescent marker in a large number of microscopic images. Assuming that each picture has the modest resolution of 1000 × 1000 pixels, the task then is, formally, to learn the mapping f: $1000 \times 1000\ R \rightarrow N$ which should yield the correct number of cells for each image, zero if there are none. To learn a mapping f defined over a small space is often feasible; but the space of all conceivable images is very large indeed, as the following parable illustrates.

Borrowing from Jorge Luis Borges (18991986) fabulous short story "The Library of Babel", let us consider The Picture Album of Babel: the album is total in the sense that it contains all conceivable images with a resolution of, say, 1 million pixels and 2563 = 16777216 color or intensity values per pixel. The album contains 167772161 000 000 which is more than 107 000 000 images. Parts of the album are rather interesting: a minuscule fraction of it is filled with images that show the evolution of our universe, at each instance in time, from every conceivable perspective, using every field of view from less than a nanometer to more than a megaparsec. Some of these images will show historic events, such as the moment when your great-grandmother first beheld your great-grandfather; while others will show timeless pieces of art, or detailed construction directions for future computing machinery or locomotion engines that really work. Other images yet will contain blueprints for time machines, or visual proofs attesting to their impossibility, or will show your great-grandmother riding a magic



(a) (b)

Figure 1.5: (a) An excerpt of a picture from The Album of Babel (b) Another picture from The Album, borrowed by contemporary artist Gerhard Richter for his stained glass design for the Cologne Cathedral, carpet, etc.

However, to us the majority by far of these images will look like the excerpt shown in Fig. 1.5, making it a little difficult to browse through it systematically in order to find the "interesting" images, such as the one showing the blueprint of the next processor generation which could earn you a lot of money. The album is so huge that, were you to order all offprints from the album at the modest size of 9 × 13cm, you would need a shoe box about 101 999 913 times the size of the visible universe to file them.

The good news is that the images of interest to us only make up a tiny part of this total album; but even so, building a general-purpose algorithm that will do well in such a high-dimensional space is hard to build.

1.4 Tasks in computer vision

So far, we have shied from defining what "computer vision" actually is. One possible taxonomy is in terms of

- image processing
- image analysis
- computer vision or image understanding / scene understanding.

All these operate on images or videos, which are usually represented in terms of arrays: a monochrome still image is represented as a two-dimensional matrix; a color or spectral image is represented as a three-dimensional array;

and a color video has two spatial dimensions, one color dimension and one temporal dimension and can hence be indexed as a four-dimensional array. In image processing, input and output arrays have the same dimensions, and often the same number of values. In image analysis, the input is an image or video and the output is a typically much smaller set of features

(such as the number of cells in an image, the position of a pedestrian, or the presence of a defect). In computer vision, the input is an image or video, while the output is a high-level semantic description or annotation, perhaps in terms of a complex ontology or even natural language. Even after five decades of research, computer vision is still in its infancy. Image processing, in contrast, is a mature subject which is treated in excellent textbooks such as [15]. The present script is mainly concerned with image analysis.

Typical tasks in image processing and analysis include:

• image restoration, to make up for deteriorations caused by deficient optics, motion blur, or suboptimal exposure

• object detection, localization and tracking

• estimation of pose, shape, geometry with applications such as human machine interaction, robotics and metrology

• estimation of motion or flow

• extraction of other features such as number of cells

• scene understanding and image interpretation as components of high-level vision, no doubt the most difficult in this list and the least developed.