Introduction to Computer Vision

Shree K. Nayar

Monograph: FPCV-0-1

Module: Introduction

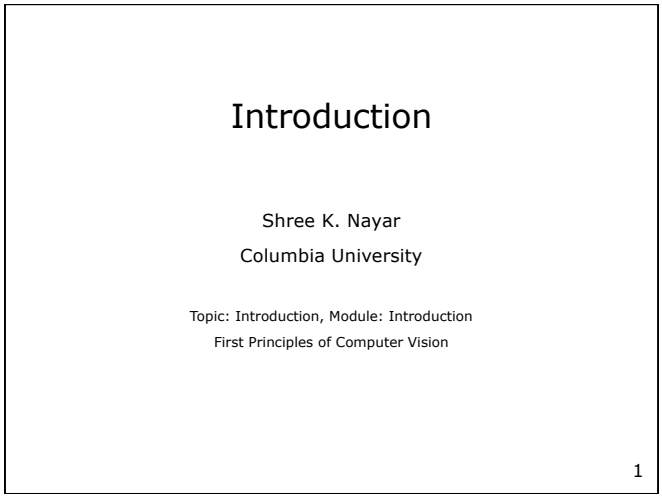
Series: First Principles of Computer Vision

Computer Science, Columbia University

February 05, 2022

FPCV Channel

FPCV Website



The poet Joseph Addison once said that "our sight is the most perfect and the most delightful of all our senses." The goal of computer vision is to build machines that can see. We have already witnessed some successful applications of vision such as face recognition and driverless cars. There is much more to come. In the next decade, we can expect computer vision to have a profound impact on the way we live our lives.

The goal of this lecture series is to cover the mathematical and physical underpinnings of computer vision. Vision deals with images. We will look at how images are formed and then develop a variety of methods for recovering information about the physical world from images. Along the way, we will show several real-world applications of vision.

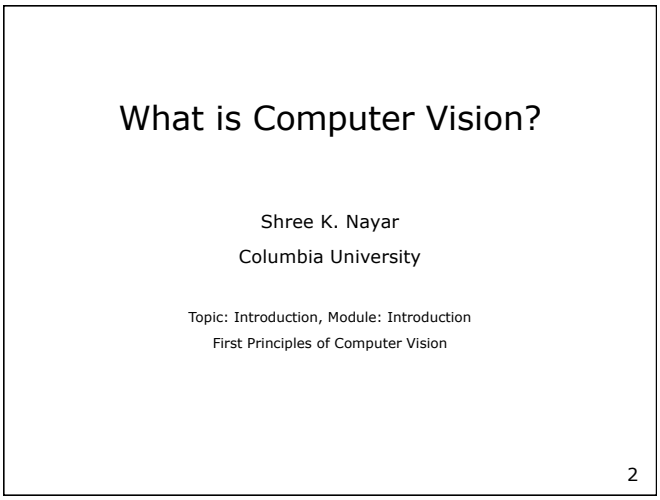
Since deep learning is popular today, you may be wondering if it is worth knowing the first principles of vision, or, for that matter, the first principles of any field. Given a task, why not just train a neural network with tons of data to solve the task? Indeed, there are applications where such an approach may suffice, but there are several reasons to embrace the basics.

First, it would be laborious and unnecessary to train a network to learn a phenomenon that can be concisely and precisely described using first principles. Second, when a network does not perform well enough, first principles are your only hope for understanding why. Third, a network that is intended to learn a complex mapping would typically require an enormous amount of training data to be collected. This can be tedious and sometimes even impractical. In such cases, models based on first principles can be used to synthesize the training data instead of collecting it. Finally, the most compelling reason to learn the first principles of any field is curiosity. What makes humans unique is our innate desire to know why things work the way they do.

I have partitioned this lecture series into 5 modules, each spanning an important aspect of computer vision. Module 1 is about imaging. Module 2 is about detecting features and boundaries. Module 3 is on 3D reconstruction from a single viewpoint. Module 4 is on 3D reconstruction using multiple viewpoints. Module 5 covers perception.

To follow any of these modules, you do not need any prior knowledge of computer vision. All you need to know are the fundamentals of linear algebra and calculus. If you happen to know a programming language, it would enable you to picture how the methods I describe can be implemented in software. In short, any science or engineering sophomore should be able to handle the material with ease.

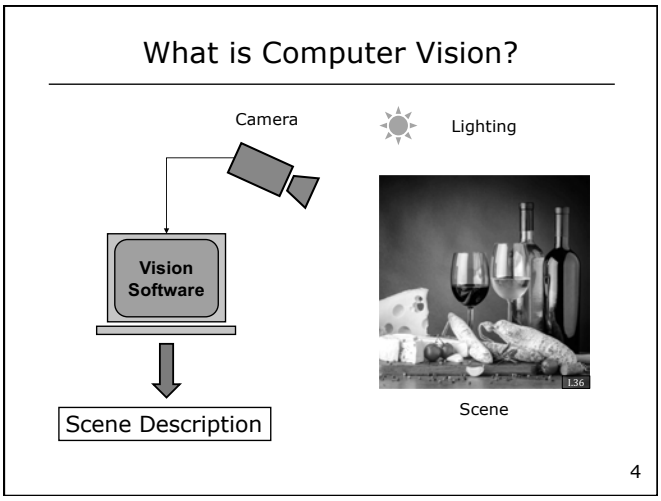
While we approach vision as an engineering discipline in this series, when appropriate, we make connections with other fields such as neuroscience, psychology, art history, and biology. I hope you enjoy the lectures and, by the end of it, I hope you will be convinced that computer vision is not only powerful but also fascinating.



Vision is our most powerful sense. It allows us to interact with the physical world without making any direct physical contact. It is believed that about 60% of the brain is, in one way or the other, involved in visual processing. Ponder that for a moment. Thanks to our vision system, we are able to effortlessly navigate through the complex world we live in and perform a variety of daily chores. In fact, our visual system is so powerful that most of the time we are unaware of how much it is doing for us.

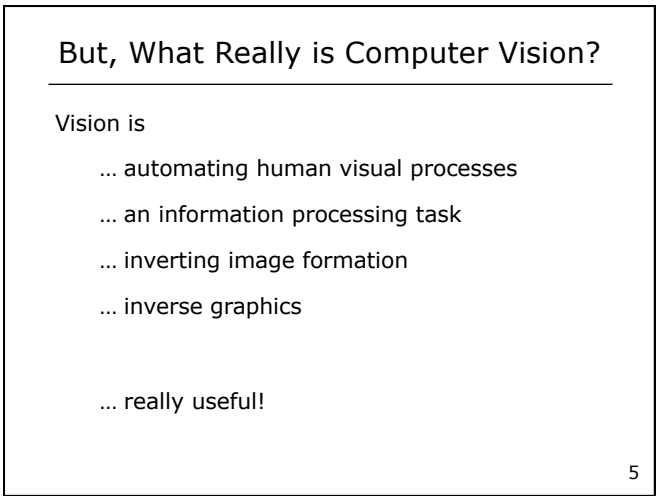
Computer vision is the enterprise of building machines that can see. You may be wondering, given that the human visual system is so powerful, why even bother to build machines that can emulate it? Well, there are several reasons. First, there are many chores we perform on a daily basis that we would rather have done by a machine so we can free up time to devote to more rewarding activities. Examples of such chores might be tidying your home and driving to work. Second, while our vision system is truly powerful, it tends to be more qualitative than quantitative. It is not particularly good at making precise measurements of things in the physical world. Lastly, and perhaps most importantly, a computer vision system can be designed to surpass the capability of human vision and extract information about the world that we simply cannot.

Here we see the basic elements of a computer vision system. On the right is a three-dimensional scene we wish to understand. This scene is lit by some form of lighting. Without light, there can be no vision. The source of lighting could be simple as in the case of a point source such as the sun, or complex as in the case of a collection of different types of lamps in an indoor setting.



The scene reflects some of the light it receives towards a camera, which plays the role of the human eye. The camera receives light from the 3D scene to form a 2D image. This image is passed on to a piece of vision software that seeks to analyze the image and come up with a symbolic description of the scene. The description could say that there are wine bottles, wine glasses, cheese, bread, and fruits in the scene. A more sophisticated vision system may be able to tell how fresh the bread is and what types of cheeses and fruits you have on the cutting board.

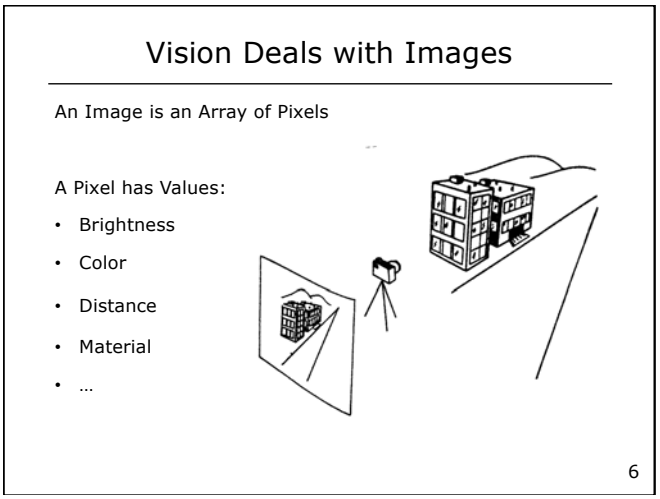
So, what would be a concise definition of computer vision? Well, it depends on the background of the person you ask. In the early years of vision, David Marr, who wrote one of the first texts on vision, defined vision as automating human visual processes. Others have viewed it more generally as an information processing task.



Berthold Horn, who wrote the book titled “Robot Vision”, viewed it as inverting image formation. An image is a mapping of the 3D world onto a 2D image. Can we now go from the 2D image back into the 3D world and say things about the objects that reside within it? Some like to view vision as the inverse of graphics. In graphics you first create detailed models of both the 3D objects in the scene and the lighting of the scene to then render a photorealistic 2D image. In vision, we are given a 2D image and wish to use it to recover the 3D models of the objects that make up the scene.

My PhD advisor Takeo Kanade used to say that, irrespective of how you define it, vision is fun! Perhaps, most importantly, vision is really useful.

Vision deals with images. An image is an array of pixels. The word pixel is short for "picture element." Typically, in an image, each pixel records the brightness and color of the corresponding point in the scene. However, pixels could be richer in terms of the information they record. For instance, a pixel could also measure the distance (or depth) of the corresponding scene point. In the future, it could also reveal the material the scene point is made of – whether it is made of plastic, metal, wood, etc.



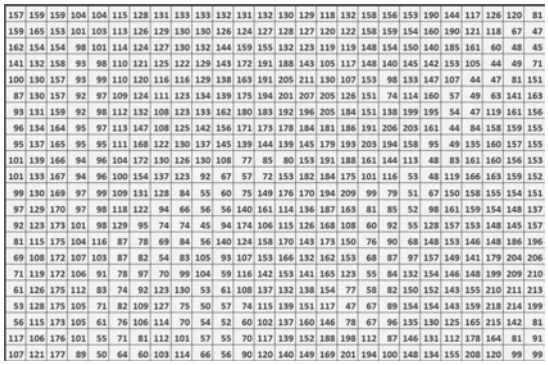
In short, with time, images will get richer in terms of the information they measure which, in turn, will lead to more powerful vision systems. One day, not too long from now, we can expect to have computer vision systems that can see things in a scene that even our powerful human visual system cannot.

We know that images are interesting. By simply opening our eyes and looking at this image, we can perceive an enormous amount of information. We are immediately able to figure out that there are two boys with one giving the other a shower, and we can perceive the three-dimensional structure of the environment, the vegetation, etc. In fact, we even get a sense of what the boy on the right must be feeling as the water falls on him and the overall playful mood of the setting – all this in a fraction of a second.

Now, in order to appreciate how difficult computer vision is, take a look at the digital equivalent of the same image shown here.

This is the array of numbers a vision system receives from the camera. It is from these numbers that we seek to extract all the information you and I perceived from the image in the previous slide. Think about that for a moment. It gives us a true appreciation for how challenging computer vision is, and that is also why it is interesting.

But When You Look Close…



8

Computer vision has been a vibrant field of research for about 50 years now. We have learned several things. First, vision is a hard problem. Second, it is a multidisciplinary field, drawing on several disciplines including optics, electrical engineering, material science, computer science, neuroscience, and even psychology. As hard as vision is, a lot of progress has been made. Today, there are many successful applications of computer vision.

Vision Research

- Vision is a Hard Problem
- Vision is Multi-Disciplinary
- Considerable Progress Has Been Made
- Many Successful Real-World Applications

9

However, there is much more to come. In the coming decades, vision is sure to play a critical role in the way we live our lives.

Let us take a look at some of the things vision is being used for today. Each one of these is a thriving industry unto itself. I should mention that these are merely examples and do not represent a complete list of vision applications.

What is Vision Used for?

Shree K. Nayar
Columbia University

Topic: Introduction, Module: Introduction
First Principles of Computer Vision

10

As you know, manufacturing is highly automated these days. Automobiles, for instance, are largely assembled by robots. Robots need computer vision to be intelligent. Without vision, robots would not be able to cope with the uncertainties that come with any real environment. For instance, if a robot is to insert a peg into a hole, it needs vision to detect any variations in the size and position of the hole. Vision-guided robotics is a major application of computer vision.

What is Vision Used For?

Factory Automation: Vision-Guided Robotics

11

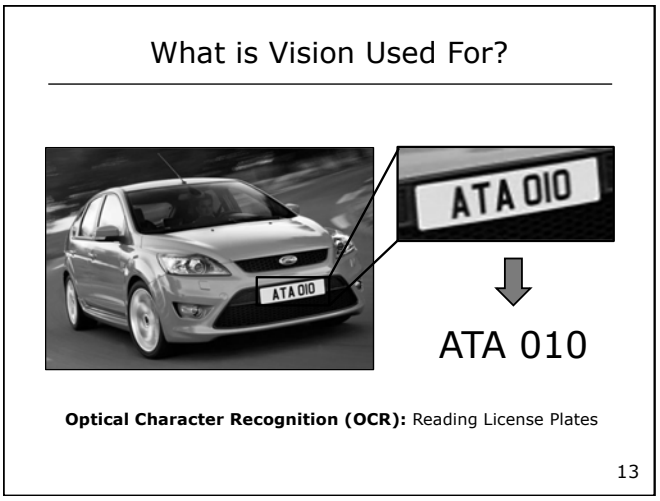
In factory automation, one of the major challenges is inspecting the quality of manufactured objects. Given the speed of manufacturing and the fact that components that go into products today can be too small for the human eye to even see, computer vision has become indispensable to modern-day manufacturing.

What is Vision Used For?

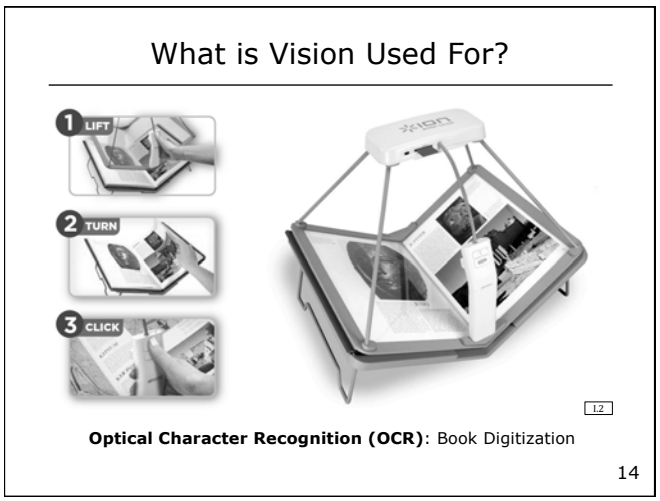
Factory Automation: Visual Inspection

12

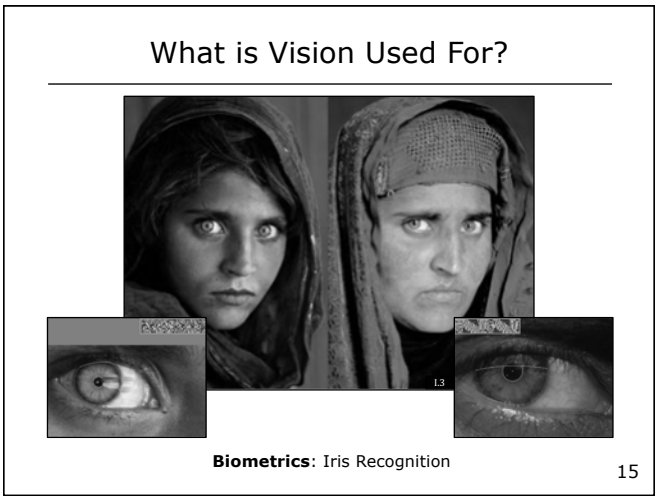
Another widely used vision technology is optical character recognition, or OCR. OCR is used today in traffic systems to identify vehicles that violate traffic rules. The license plates of these vehicles are automatically read, and tickets are mailed to violator's home.



Character recognition, as you can imagine, has many other important applications, such as digitization of physical books, authentication of signatures on checks, and reading mailing addresses on envelopes and packages received by postal services. OCR is now even available in phone apps that enable you to translate in real time a sign in one language into a language you understand.



Vision plays a critical role in the field of biometrics, where one's physical characteristics are used to determine their identity. Iris recognition is a widely used biometric today. Take a look at the high-resolution images of the eyes of these two people. It turns out that the intricate patterns seen in a person's iris are unique to them, almost as unique as their DNA, and can be used to determine their identity with very high confidence.



A vision technology that is ubiquitous today is face detection. It is considered to be one of the most successful applications of vision. Faces can be robustly detected in images under different poses and illuminations. Face analysis can also be used to recognize the identity of a person in an image. This technology has far-reaching implications. It can be used to organize photos in your personal collection and to find suspicious persons in security applications.

What is Vision Used For?

Biometrics: Face Detection and Recognition

16

A recent and interesting application of vision is intelligent marketing. Here, you see a vending machine I have used in Shinagawa Station in Tokyo. As a person approaches the machine, it identifies the gender and rough age of the person and displays products that are most likely to be of interest to the person.

What is Vision Used For?

Intelligent Marketing: Vending Machine with Face Detection

17

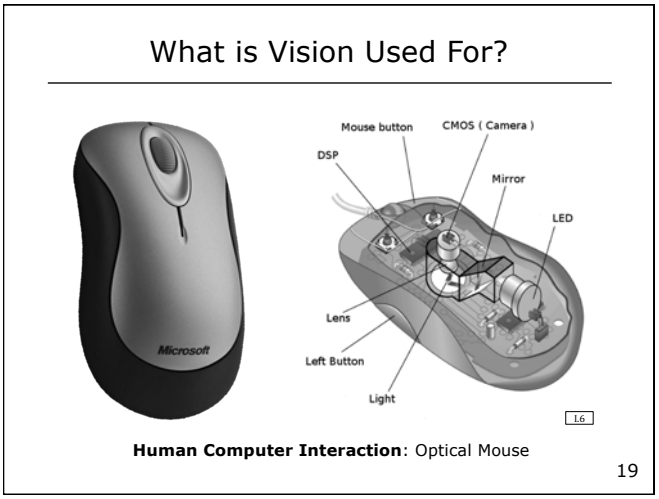
Modern vision algorithms can also robustly track people moving through space. In the context of surveillance, this technology is used to follow a person as they move through a crowd, even as they get obstructed by other people or objects in the scene. In fact, when a person leaves the field of view of one camera, they can be handed off to another camera that continues to track them.

What is Vision Used For?

Security: Object Detection and Tracking

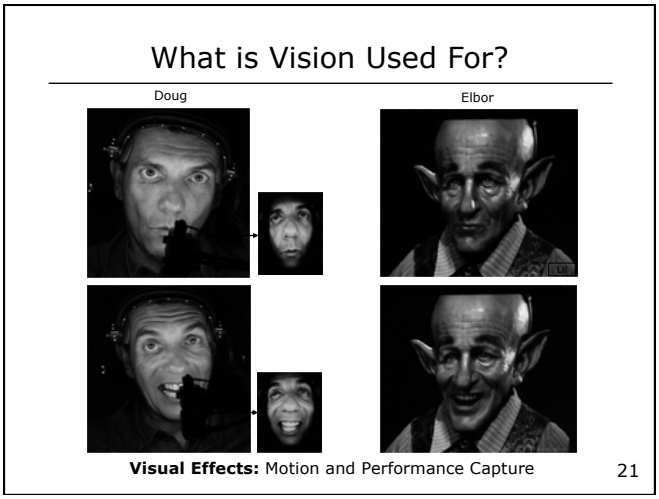
18

One application of vision that is ubiquitous, but one that you may be unaware of, is the optical mouse. Inside the mouse is a complete vision system that tracks the movement of the pattern, or texture, of the surface on which the mouse sits. This is done using a low-resolution camera with a very high frame-rate, enabling the mouse to precisely estimate its motion with respect to the surface it sits on. This information is used by your computer to control the position of the cursor.



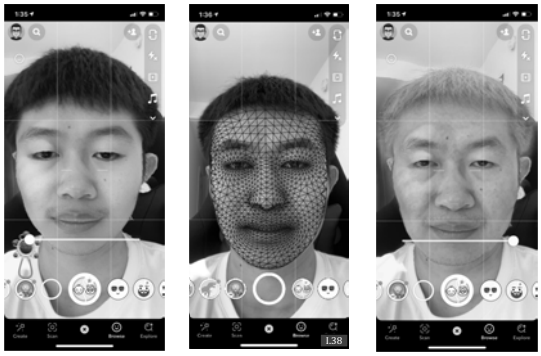
Another popular application of vision is in gaming consoles. Here you see a player using Microsoft's Kinect. Kinect is packed with vision technology, enabling it to capture the motion of the player's full body. This has given rise to a new breed of engaging interactive games.

Vision plays a vital role in creating special effects in movies and animations. Here, you see Doug Roble on the left with a camera attached to his head that is watching him. The expression on Doug's face is computed in real time and transferred to the face of Elbor, who is a virtual character. Using this technology, Doug Roble recently gave an entire TED talk via a virtual character.



Vision is also being used to create augmented reality (AR) technologies. Here, you see Jian Wang's face captured using the Snapchat camera. Jian's 3D face model (middle) is computed in real time. This model is used to modify the appearance of Jian's face. For instance, you can see what Jian might have looked like when he was a young boy (left) or what he might look like when he becomes an old man (right).

What is Vision Used For?



Augmented Reality: Face Manipulation

22

A useful application of vision is landmark recognition. This is within the realm of what is called visual search. By simply taking a picture of any well-known landmark, such as this monument, one can immediately get detailed historical information regarding the monument.

What is Vision Used For?

Visual Search: Landmark Recognition

23

Vision is not just useful but essential in certain fields such as space exploration. Here, the Mars rover uses an array of cameras to extract and send detailed information about the terrain of Mars back to Earth. This is a scenario where vision is the only way humans can explore a region that is inaccessible to them.

What is Vision Used For?

Autonomous Navigation: Space Exploration

24

One application of vision that is widely talked about today is the driverless car. These cars use a wide range of cameras – visible light, infra-red, and depth – to measure their surroundings with high precision and detail. This information is used by algorithms to enable the car to make decisions in a variety of driving scenarios. There is little doubt that driverless cars will soon become a part of our everyday lives. This would not be possible without the advances made by computer vision.

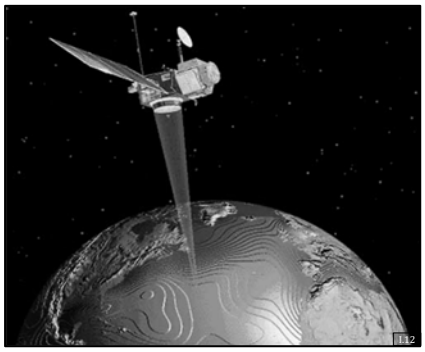
What is Vision Used For?

Autonomous Navigation: Driverless Car

25

Another area that exploits vision is remote sensing. Here, you see a satellite orbiting the Earth. High-resolution cameras on the satellite are used to create 3D maps of the Earth's surface, monitor natural disasters, surveil enemy territory during war, and track the effects of climate change on the planet.

What is Vision Used For?



Remote Sensing

26

An important application domain for vision is medical imaging. While vision is most often applied to visible light images captured by consumer cameras, the algorithms developed for such images can be modified and used to analyze medical images such as X-ray, ultrasound and magnetic resonance images. Here you see a magnetic resonance image (MRI) from which anatomical structures can be automatically detected and analyzed to help diagnose the patient.

What is Vision Used For?

Medical Image Analysis

27

I hope I have convinced you that vision has enormous utility. That is why it is a topic of wide interest today.

Before we begin to develop tools to help us solve vision problems, it is worth taking a quick look at how our human visual system works.

How do Humans do it?

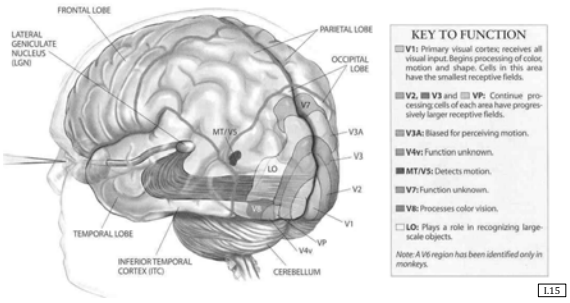
Shree K. Nayar
Columbia University

Topic: Introduction, Module: Introduction
First Principles of Computer Vision

28

As shown here, images of the world are captured by each of our two eyes. Some early visual processing takes places in the eye itself. That is, the information recorded by the retina is processed by cells in the retina to reduce the information that needs to be transmitted to the brain. This information travels via the optic nerve to the lateral geniculate nucleus where it is relayed to the visual cortex – the part of the brain in the back that performs most of the visual processing.

Human Eye and Visual Cortex



1.15

Vision is easy for us

But, we don't fully understand how we do it!

29

As you can see in this map of the cortex, there are regions of the cortex that perform different functions such as the perception of shape, color, motion, etc. While we know roughly where each type of analysis takes place and roughly how many neurons there are in each of these regions of the cortex, we are very far from having a detailed architecture, or "circuit diagram", of the human visual system. In short, we do not know enough about the visual cortex to replicate it using electronics.

So, what do we do? We reinvent. This might sound unfortunate to you, but not quite. As you can imagine, there are many applications of vision that require functionality and precision that go well beyond what the human visual system is capable of. While human vision is remarkable in its versatility and is able to cope with many complex real-world situations, it is more of a qualitative system than a quantitative one.

For instance, if you wanted to know how many millimeters long a pen is, the human visual system can only give very rough estimates. Such estimates are not useful in domains such as factory automation, medical imaging, or autonomous navigation of robots and cars. While no computer vision system has yet been developed that is as versatile as the human one, there are many computer vision systems in use today that demonstrate much higher precision and reliability than ours. In short, for many tasks that require vision, human vision may indeed be the wrong system to emulate.

What do we do?

Reinvent!

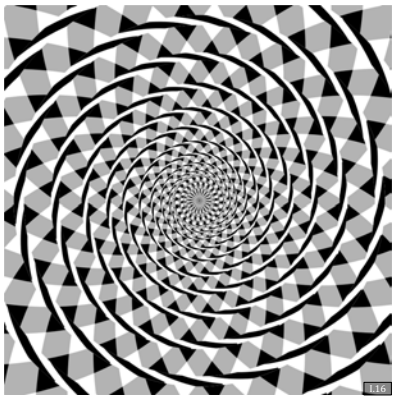
30

Furthermore, human vision is more fallible than we may like to believe. When you and I perceive something incorrectly, we do not have a voice in our head telling us we are wrong. We see what we see and believe it to be accurate.

To demonstrate this, let us take a look at some well-known optical illusions.

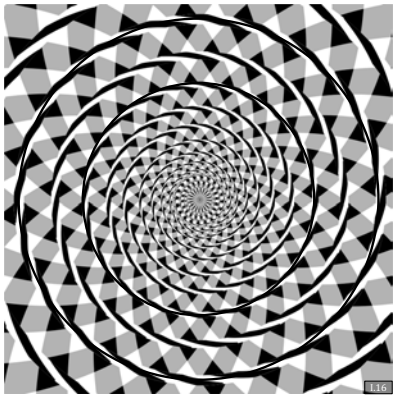
On the left is Fraser's spiral. You should be seeing a spiral emerging from the center of the image. Well, it turns out there is no spiral in the photo. You can verify this by following any one of the curves – you will see that you end up where you started. That is, Fraser's spiral is actually a set of perfectly concentric circles as shown on the right.

Illusions: Fraser's Spiral



31

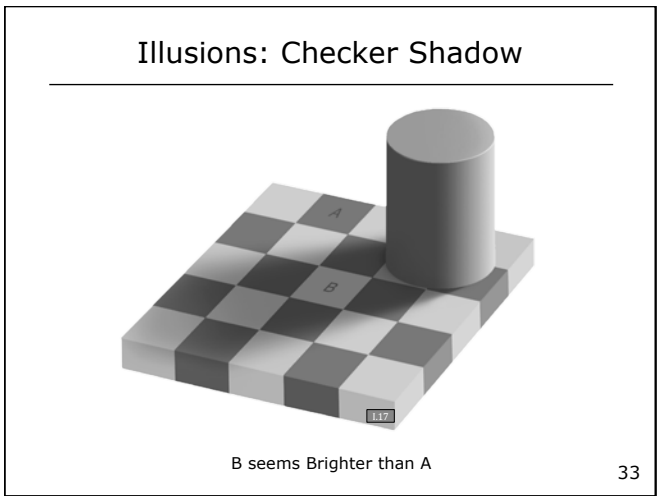
Illusions: Fraser's Spiral

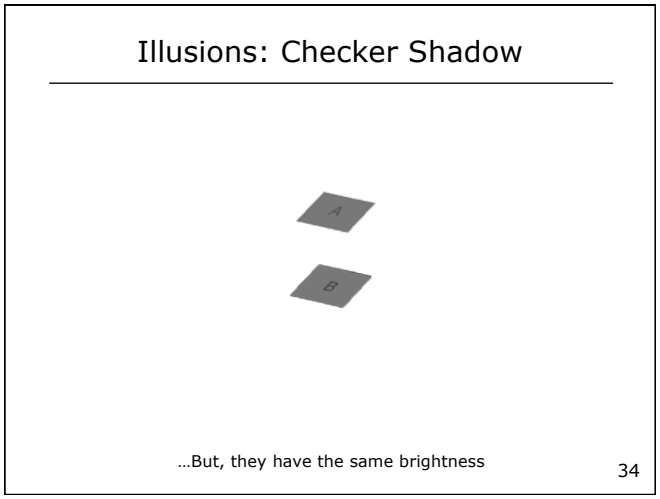


32

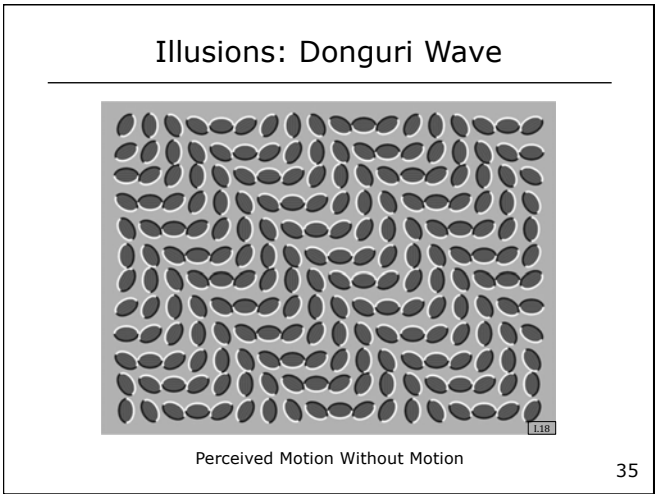
Here, on the left, is the checker shadow illusion from Ted Adelson. We would all agree that patch A is made of darker material than patch B. It turns out that we perceive this to be the case because our visual system is able to determine that the illumination is varying over the scene. We first estimate this spatially varying illumination and then use it to compensate for the brightness at each point in the scene. The end result is that we perceive patch A to be of lower reflectance (darker material) than patch B.

Now, if you look at patches A and B in isolation, that is, if you cut them out of the rest of the image as done on the right, you see that they have exactly the same brightness.





Here is the Donguri wave illusion. It is a single image and not a video. Yet, when you move your eyes around the image, you perceive the leaves to be moving around. Clearly, an illusion.



Here is an example of forced perspective. The person standing on the right seems to be much shorter than the one on the left. In fact, they are almost exactly the same height. The room itself is not a cuboid but rather tapered such that the distance between the floor and the ceiling increases with distance from the camera. This changes your perception of the relative sizes of objects in the room.

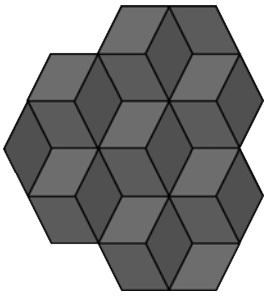
Illusions: Forced Perspective

These two people are of the same height!

36

Then, there are visual ambiguities. These are not really illusions. It is just that the image itself, since it is 2D while the world is 3D, can lead to multiple interpretations of objects in the scene. In this case, if you stare at one of the vertices, you can lead yourself to believe that it is a corner that is convex (cube popping out) or concave (cube pushed in). In fact, if you keep staring at the vertex, you will find yourself flipping between these two interpretations.

Visual Ambiguities



Six Cubes or Seven Cubes?

37

There are even higher levels of ambiguity. The image on the left can be perceived to be either of a young girl turned away from the viewer or the profile of an old woman with a large nose.

On the right, it could be a vase or two heads facing each other.

Visual Ambiguities

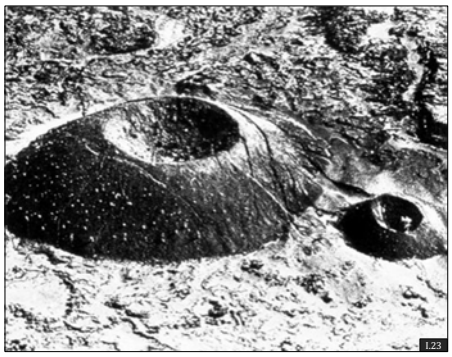
Young-Girl/Old-Woman

Face/Vase

38

Here is my favorite ambiguity. On the left you see a large mound with a small crater in the center. Now, what would you see if you turned this image upside down? You would expect to see the large mound hanging upside down, right? As seen on the right, in fact, you see a large crater with a small mound in the center. You perceive this because we live in a world where the light is expected to come from above – from the sun for instance. We therefore invoke this assumption – that the illumination is from above – while interpreting the shape of an object from its shading.

Visual Ambiguities



Crater on a Mound

39

Visual Ambiguities

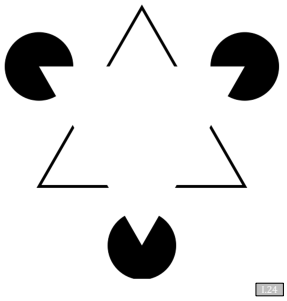
Mound in a Crater

40

Finally, there is the famous Kanizsa triangle. Here we perceive a white inverted triangle in the center. In fact, the triangle appears to be slightly brighter than the white background. It also appears to be closer in depth, seeming to sit above the rest of the scene.

Of course, there is no triangle here. It is just three “pac-man” like fragmented discs that are precisely aligned to give the illusion of a triangle.

Seeing vs. Thinking



Kanizsa Triangle

41

This example tells us that there is seeing and then there is perceiving. Our eyes see the three “pac-man” discs, but from their arrangement our brain infers the existence of a triangle.

Now, let us take a quick look at the topics covered in this lecture series.

Topics Covered

Shree K. Nayar
Columbia University

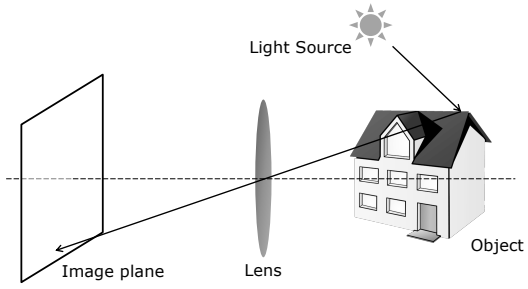
Topic: Introduction, Module: Introduction
First Principles of Computer Vision

42

We start with image formation, where we look at how the 3D world is projected by a lens to form a 2D image. We would like to understand the geometric and photometric relation between a point in 3D and its 2D projection in the image.

Image Formation and Optics

Where do Images Come From?



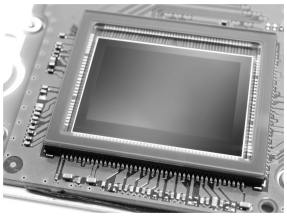
Projection of 3D world on a 2D Plane

43

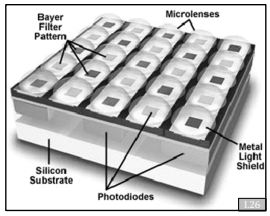
Next, we look at image sensors that are used to convert the optical image formed by a lens into a digital image. Image sensor technology has evolved rapidly, enabling us to capture digital images that today exceed what film could do in the past. If there is one reason for the imaging revolution we have witnessed in the past decade or so, it is the remarkable improvements made in image sensor technology.

Image Sensors

Convert Optical Images to Electrical Signals



Consumer Image Sensor



Typical Structure of Image Sensor

44

The simplest type of image is the binary image, which is a two-valued image (right) obtained by simply thresholding a captured image (left). You end up with white (or 1) for object and black (or 0) for background. Such images are often used in factory automation. They are easy to store and process. We look at how they can be used to solve simple vision problems.

Binary Images

Two-Valued Images: Easy to Store and Process

Grayscale Image

Binary Image

45

Next, we will devote two lectures to image processing, which seeks to transform a captured image into a new one that is cleaner in terms of the information we want to extract. In this example, you can see a noisy captured image on the left that is processed to create the one on the right. In this processed image, the noise is more or less removed while the edges are preserved.

Image Processing

Transform Image to New One that is More Useful

Input Image

Edge-Preserved Smoothing

46

With image processing tools under our belt, we are in a position to extract useful features from images. From the perspective of information theory, edges in an image are of great importance. We start by developing a theory of edge detection. Based on this theory, we develop a few different edge and corner detectors.

Edge and Corner Detection

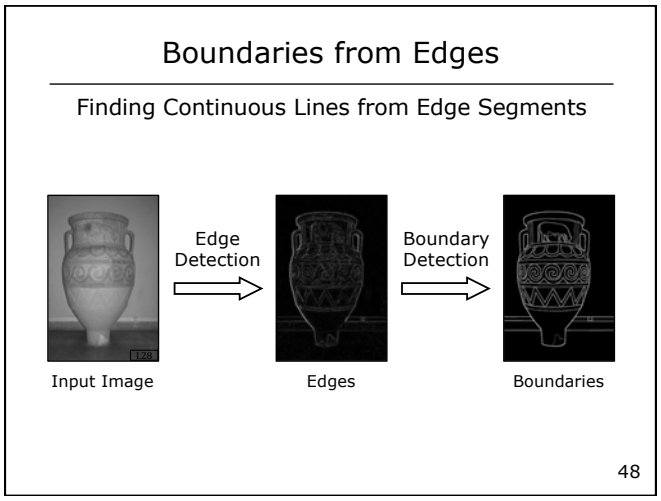
Detecting Intensity Changes in the Image

Input Image

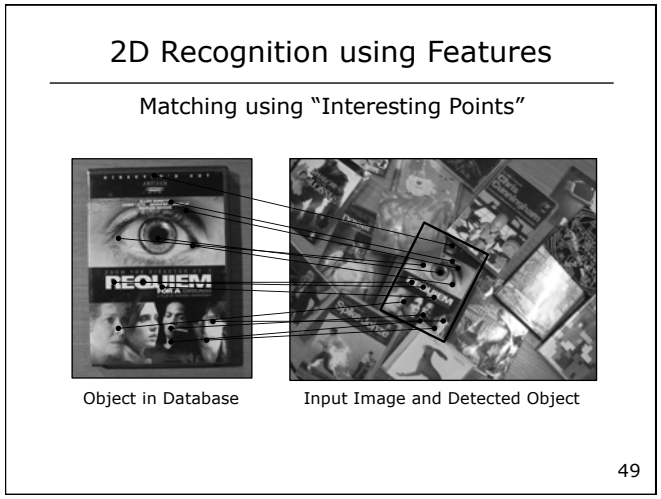
Edges

47

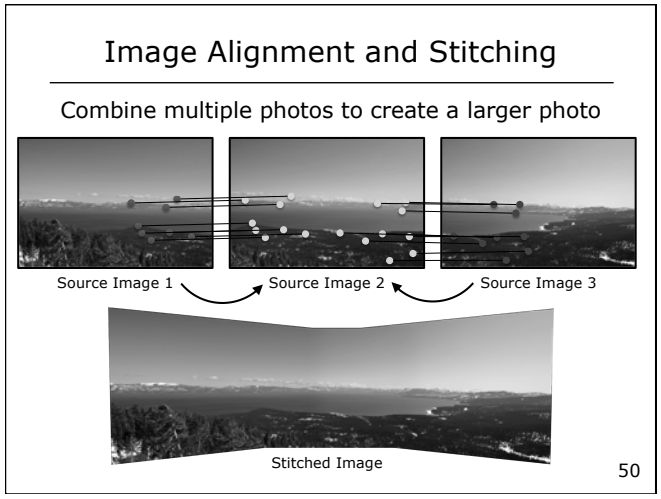
When you apply an edge detector to a typical image, you get a lot of edges, but they are not related to each other. In order to extract objects from an image, we need to go from edges to boundaries. When we look at the edge image in the center, we can quickly group edges that belong to the same boundary or contour. It turns out that this grouping process is not as easy as it seems. We develop a variety of algorithms that can group edges to extract boundaries.



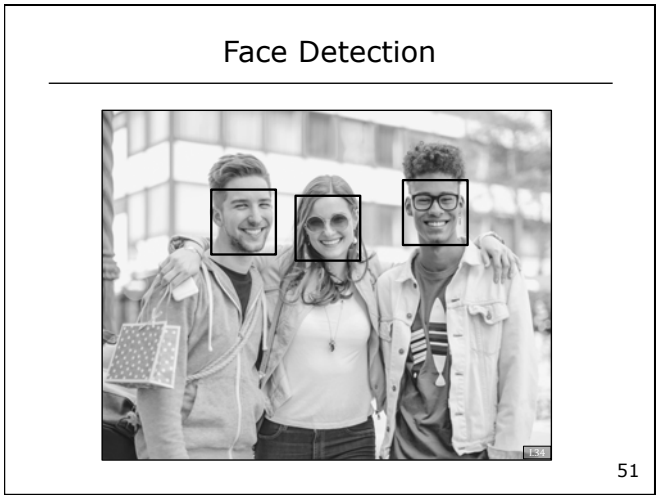
One feature detector we describe in detail is the Scale Invariant Feature Transform (SIFT), which can detect interesting blobs in an image. SIFT features can be used to robustly detect and recognize planar objects in an image, even when they are scaled, rotated, and occluded by other objects.



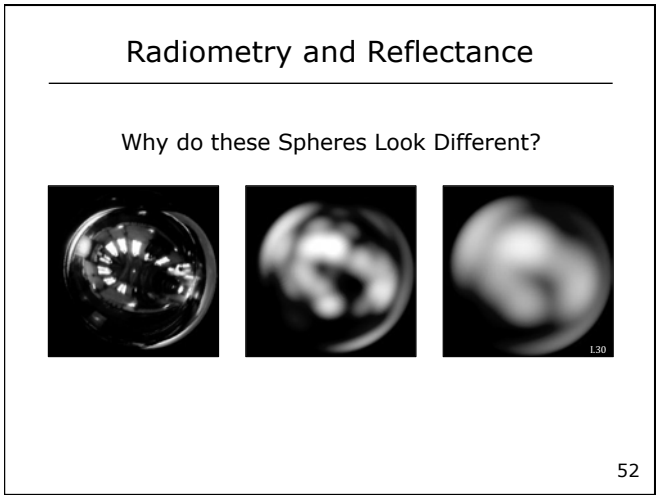
As an example application of feature detection, we develop an algorithm that can take a set of overlapping images of a scene taken from roughly the same viewpoint (top row) as input and produce a single seamless panorama (bottom). This method, called image stitching, is available on most smartphones.



Next, we discuss the widely used technique of face detection. We develop an algorithm that can efficiently and robustly find faces in an image. One of the key challenges here is to be able to reliably discriminate between faces and non-faces and handle faces with different skin tones, expressions, illuminations, and poses.

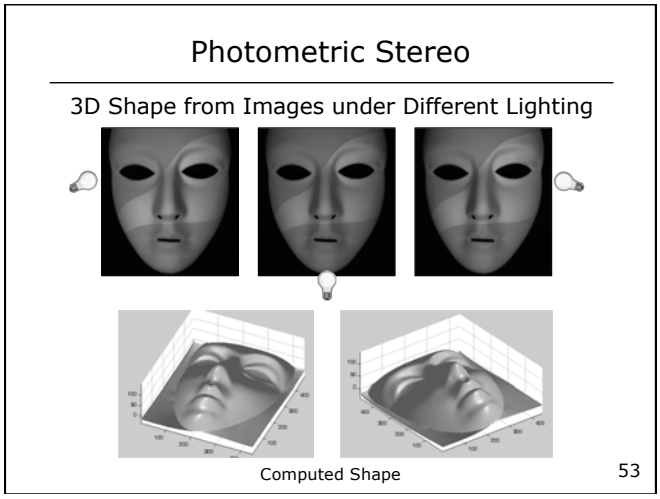


Everything we have discussed thus far focuses on extracting information in image coordinates, that is, in two dimensions (2D). Next, we lay the groundwork for developing algorithms that recover the three-dimensional (3D) structure of a scene from one or more images.

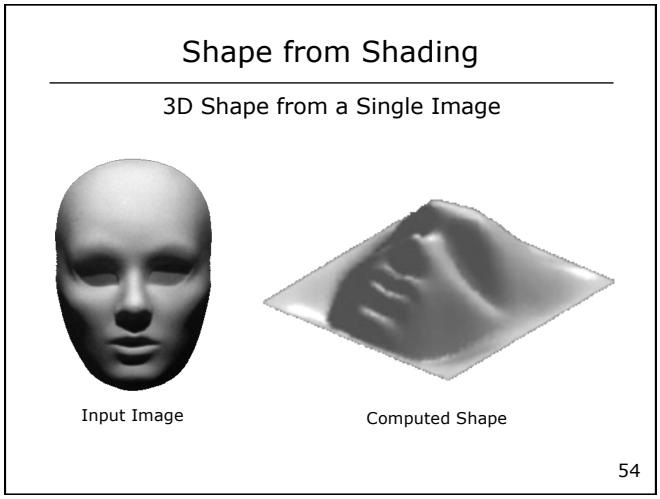


The first topic in this context is radiometry and reflectance. We begin by defining the fundamental concepts of radiometry, which has to do with measuring light. We will establish a relation between the brightness of a point in the scene to its brightness in the image. Then, we explore why different materials appear the way they do. We discuss a small number of popular reflectance models that can each describe a wide range of materials found in the real world.

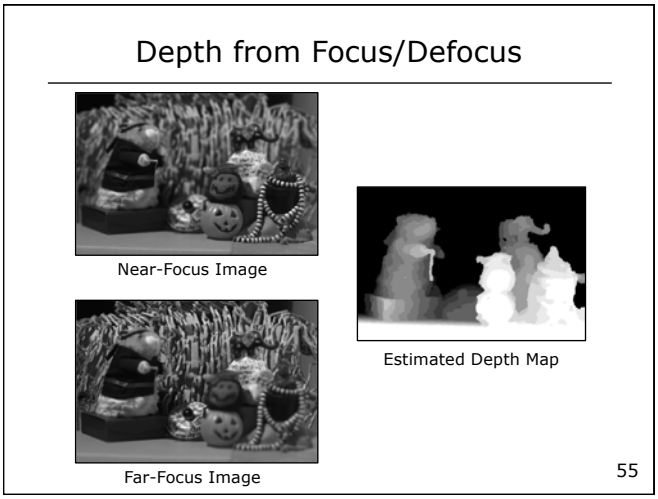
It turns out that if we take a few images of an object under different, known lighting conditions, we can compute the surface normal at each point on the object. This method is called photometric stereo. In the case of a continuous surface, the measured surface normals can be used to reconstruct the shape of the object.



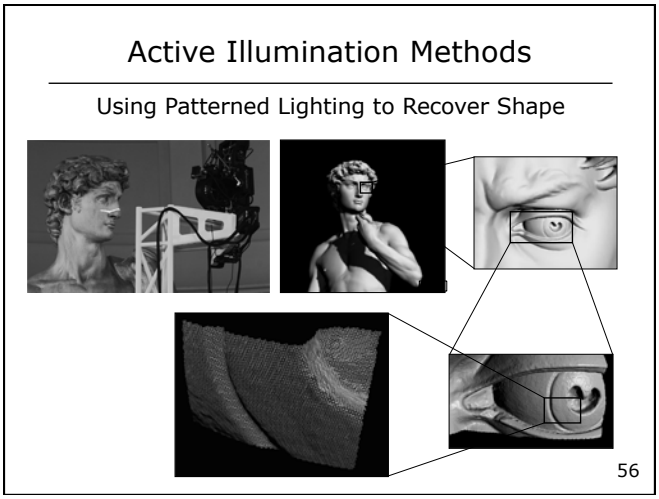
Next, we look at shape from shading, a more challenging problem, where we seek to extract the 3D shape of a surface from a single shaded image. We use experiments conducted with humans to show that shape from shading is an under-constrained problem. By using a few assumptions, we show that shape can indeed be recovered from a single image using shading.



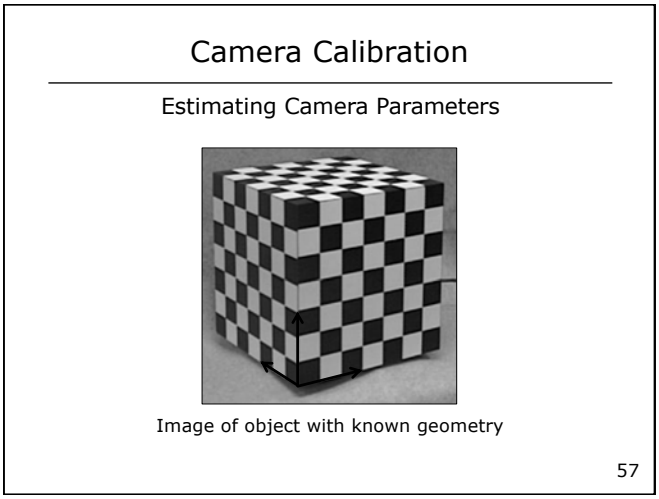
As you know, when you vary the focus setting of your camera, you see objects in the scene come into and go out of focus. In this sequence of images, exactly when a point comes into focus is a function of its depth from the camera. We develop algorithms for recovering the 3D structure of a scene using both focus and defocus.



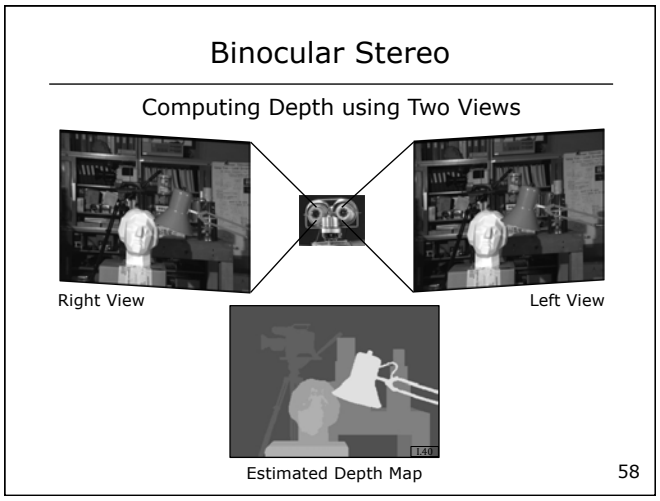
Our next topic is active illumination. In many real-world applications of vision, such as factory automation, we have the ability to control the illumination of objects being imaged. When this is possible, we can develop very efficient and accurate methods for recovering the shapes of objects. In this example, we see the statue of Michelangelo's David, which has been scanned using active illumination to produce a remarkably accurate 3D model.



If we wish to precisely measure the height of a bottle in millimeters by taking images of it, we need to first know how the position of a point in an image, which is measured in pixels, relates to the position of the corresponding point in the 3D world. Relating 2D image coordinates to 3D world coordinates requires us to know the various parameters related to image formation. The process of estimating these parameters is called camera calibration. We show how a single image of a 3D object of known dimensions, such as the cube here, is enough to compute all the parameters of the camera.



Next, we present binocular stereo. Try this simple experiment. Make the index fingers of both your hands point out while folding your other fingers in. Now, hold your left hand out in front of you and your right hand behind you. Then, shut any one of your two eyes and bring your right hand from behind quickly to make both of your index fingers touch. Not easy, right? This is because with just one eye, you lose quite a bit of your ability to perceive depth.

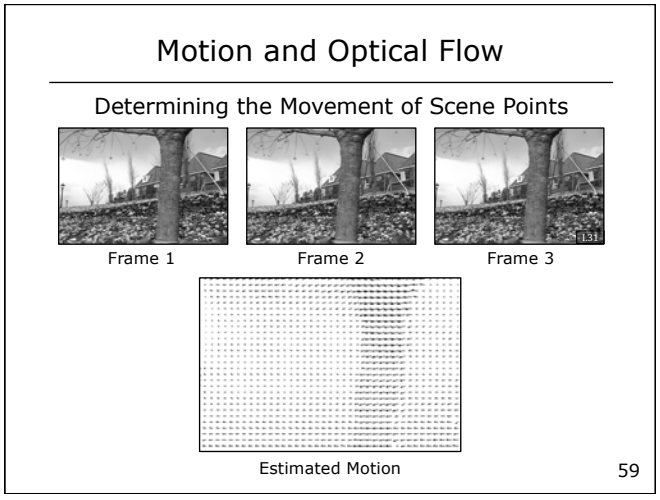


Now, repeat the same with both of your eyes open. You should find it much easier to make the two fingers meet.

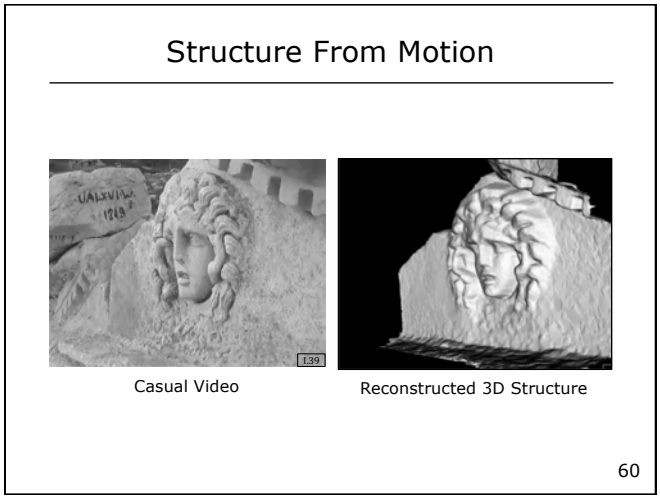
We use two eyes to help us perceive the 3D structure of the world in front of us. The two eyes capture two slightly different views of the world in front of us. The small differences in these two views are used to estimate the depth of each point in the scene.

We will describe algorithms for achieving the same using two cameras. Here you see a right camera image and a left camera image and the depth of the scene computed from them. In the depth image, the closer a point is, the brighter it is.

Thus far, we have assumed the world to be static while we capture our images. We know that, in reality, everything is in motion all the time. We look at how the motion of a point in 3D relates to its motion in the image. The visible motion of a point in an image is called optical flow. Based on first principles, we develop an algorithm for estimating optical flow (bottom) from a sequence of images captured in quick succession (top).

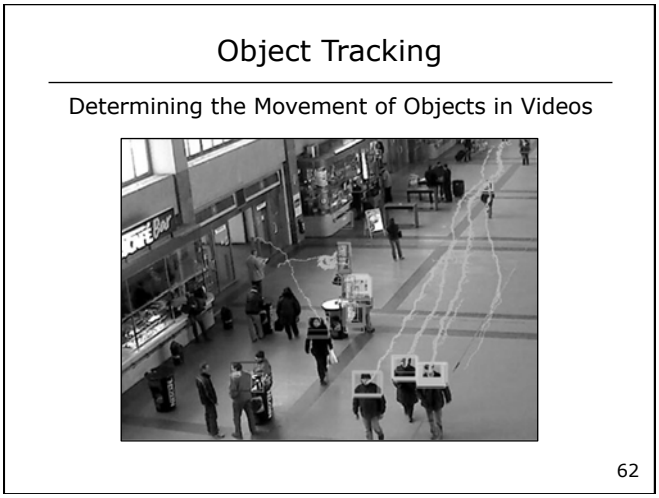


Imagine capturing a video of a structure such as the sculpture shown here by simply moving a camera around it. Note that this is a casually captured video, and we do not know the motion of the camera in 3D. It turns out that, even without knowing the motion of the camera, we can compute the 3D structure of the scene (right). Interestingly, in addition to the structure of the scene, we can also determine the motion of the camera. This method is called structure from motion.

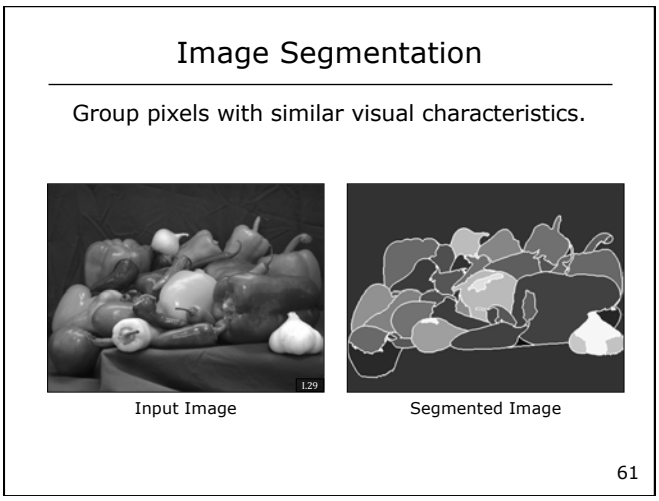


In the remaining lectures, we will cover topics related to higher levels of visual perception.

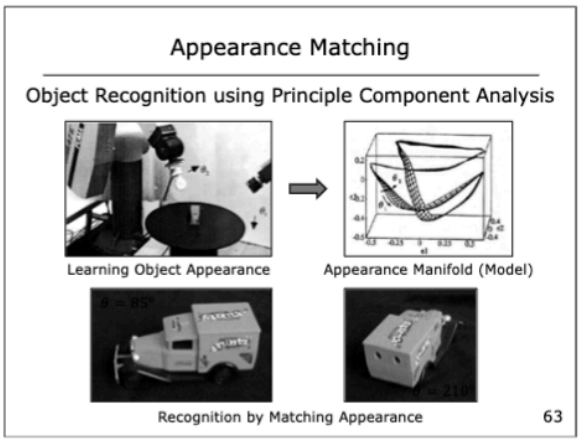
We start with the problem of tracking objects as they move through 3D space. In this example, we see that the system is able to produce a separate track for each person walking through a public space. Note that such an algorithm must be able to handle objects that briefly go out of view when they are obstructed by other objects.



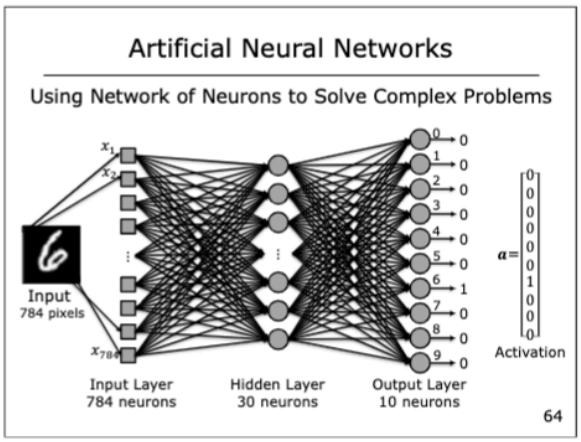
Next, we look at the important problem of image segmentation. We develop algorithms that can take an image (left) as input and segment it into clearly defined regions (right), where each region more or less corresponds to a single physical object. Segmentation is an ill-defined task since what exactly constitutes an object often depends on the context. We develop a few different approaches to image segmentation.



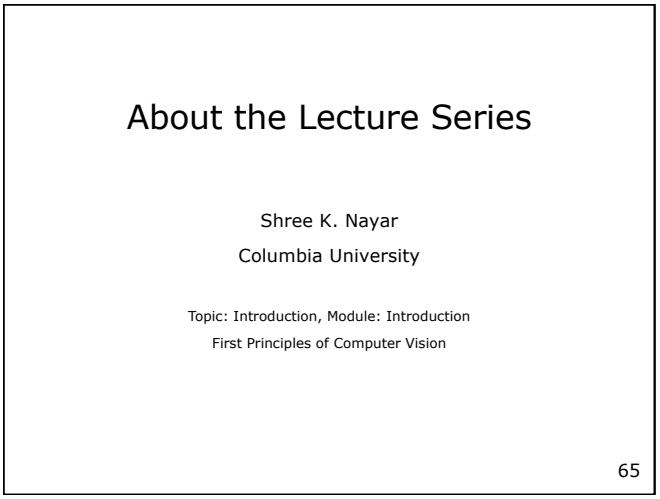
The last problem we look at is recognition. The first approach to recognition we discuss is called appearance matching. We capture many images of an object under different poses and illumination conditions. Dimension reduction is used to compute a compact appearance model of the object from its images. When the object appears in a new image, it is segmented and recognized using its appearance model. The model also reveals the pose and lighting of the object.



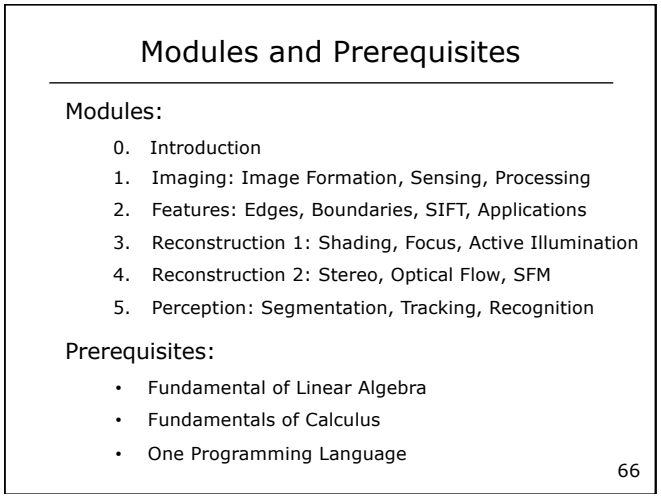
Finally, we discuss the use of artificial neural networks for solving complex recognition tasks. We first describe the concept of a neuron and how a network is constructed using a large number of neurons. We then show how a network can be efficiently trained using the back-propagation algorithm. We conclude with a few examples of the use of networks for solving challenging visual recognition problems.



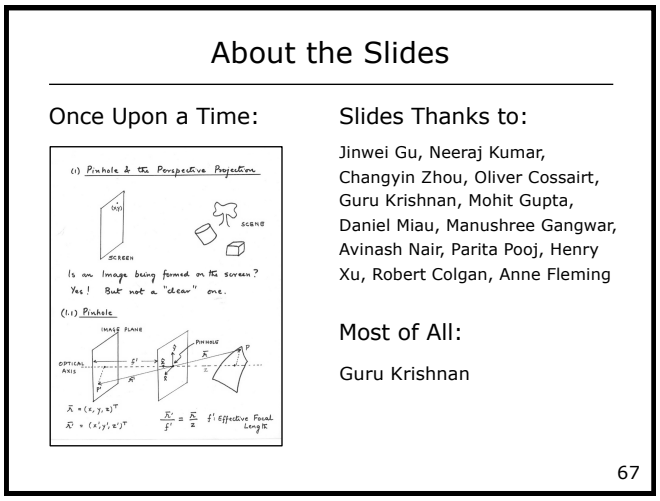
The lecture series is organized as 5 modules in addition to this introduction. Module 1 focuses on imaging and will cover image formation, image sensing and image processing. Module 2 is about features and boundaries and will include edge detection, boundary detection, the SIFT feature detection, and applications like stitching panoramas and detecting faces. Next, in Module 3, we explore ways of reconstructing a 3D scene using one or more images taken from a single viewpoint. This module includes photometric stereo, shape from shading, depth from defocus, and active illumination methods. In Module 4, we explore reconstruction methods that use images taken from multiple viewpoints. These methods include binocular stereo, optical flow, and structure from motion. Finally, in Module 5, we discuss perception, which includes higher levels of visual processing such as tracking, segmentation, and object recognition.



To follow these lectures, you do not need any prior knowledge of computer vision. All you need to know are the fundamentals of linear algebra and calculus. If you happen to know a programming language, you would be able to imagine how the methods we discuss can be implemented in software.



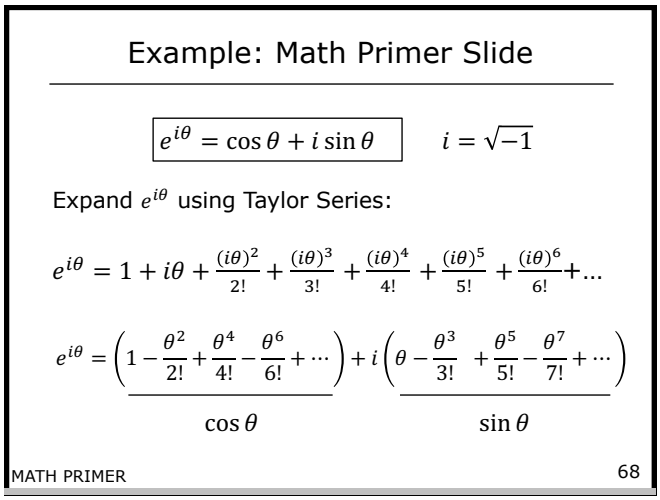
Now, a few words about the slides I use in the lectures. I have been teaching a course on computer vision at Columbia for many years. In the initial years, before PowerPoint and Keynote became popular, I used hand-made overhead projector slides like the one shown here. In recent years, several of my students and postdocs have helped me create the slides I am now using. Also, new content has been added to the lectures as the field has evolved.

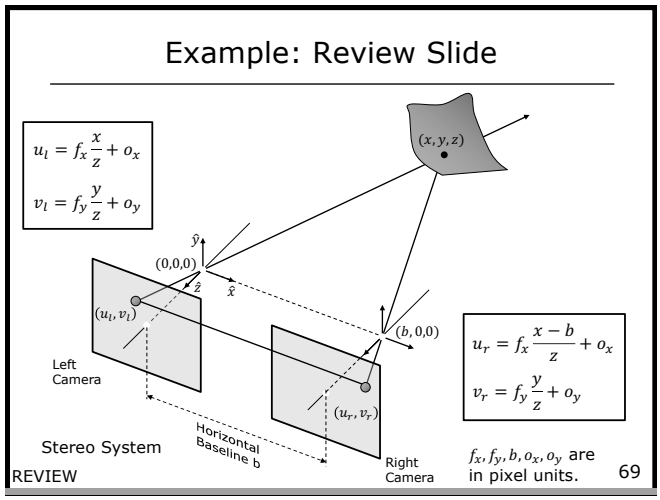


Many of my students and postdocs have contributed in small and big ways to these slides. In particular, I would like to thank Jinwei Gu, Neeraj Kumar, Changyin Zhou, Oliver Cossairt, Guru Krishnan, Mohit Gupta, Daniel Miau, Avinash Nair, Parita Pooj, Henry Xu, Robert Colgan and Anne Fleming for their efforts. Most of all, I would like to thank Guru Krishnan who did the bulk of the work. Without Guru's efforts, I am not sure I would have been able to create this lecture series.

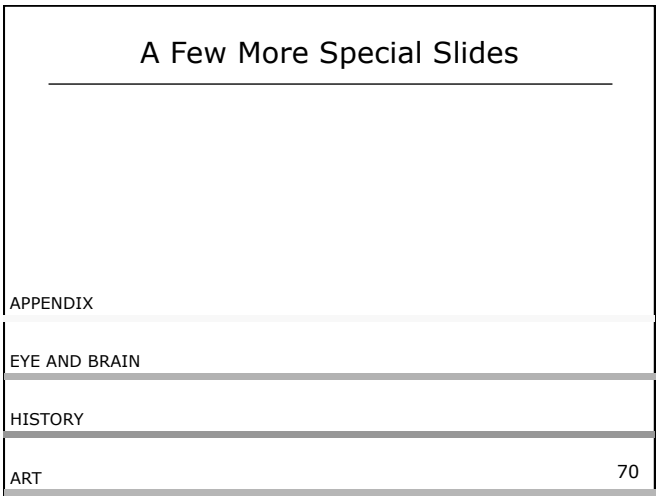
The slides come in a few different flavors. The one on the left is labeled at the bottom as a math primer. I have several of these in the lectures. Each math primer highlights a mathematical concept that is not only needed for vision, but one that I believe is useful to any engineering or computer science student.

Every now and then, you will see a review slide like the one on the right that is used to recap a concept that was covered in an earlier lecture.

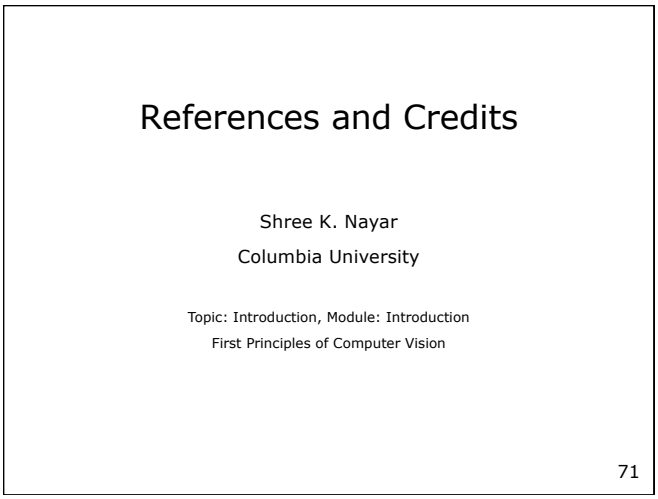




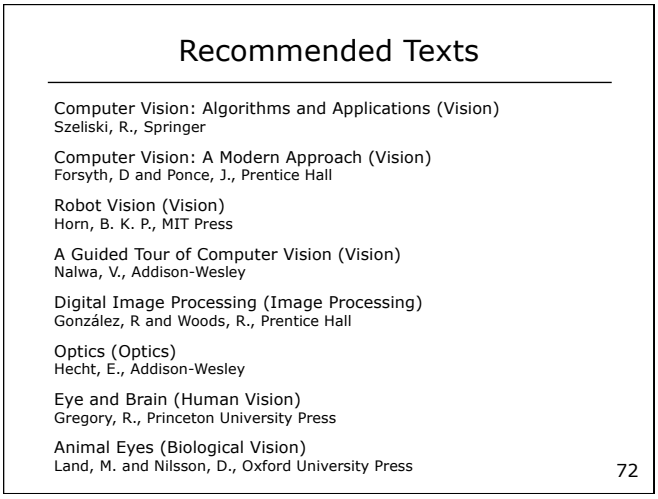
There a few other categories of slides you will see along the way. These include appendices, eye and brain, history, and art. While computer vision has historically been considered to be a subfield of computer science, it has strong connections to many other fields. I have always found these connections interesting and so have highlighted them when appropriate.



The material covered in these lectures comes from varied sources. I have used published papers and textbooks.



Here are a few of the texts I have used. The one that overlaps most substantially with these lectures is “Computer Vision: Algorithms and Applications” by Rick Szeliski. “Computer Vision: A Modern Approach” by Forsyth and Ponce is also a book I would encourage you to refer to. When it comes to first principles of many vision topics, it is hard to beat “Robot Vision” by Berthold Horn. Few authors are quite as precise as Horn. For a concise and well-written overview of vision, I recommend Vic Nalwa’s “Guided Tour of Computer Vision”.



We have two lectures on image processing. There are many excellent books on the topic and if you needed to pick one, I would suggest “Digital Image Processing” by Gonzalez and Woods. For all things related to optics, I strongly recommend the classic by Hecht. There is a small but wonderful book called “Eye and Brain” by Richard Gregory, which has a lot of great nuggets related to human vision. Finally, if you are interested in biological eyes of all kinds, I strongly recommend “Animal Eyes” by Land and Nilsson. It is a thin book but densely packed with information.

Finally, I believe any lecture related to vision must be visual. Else, it is a lost opportunity. I have used many visuals from varied sources in my lectures and at the end of each lecture you can find the credits related to the photos and videos I have used.

Image Credits

I.1 https://www.automation.com/images/article/omron/MVWP3.jpg
I.2 Adapted from ION Sound Experience. http://www.ionaudio.com/downloads/booksaver_2011_overview.pdf
I.3 Steve McCurry. Used with permission.
I.4 Anton Milan. Used with permission.
I.5 http://www.designboom.com/design/acure-digital-vending-machine/
I.6 http://howthingswork.org/electronics-how-an-optical-mouse-works/
I.7 Oliver Berg dpa.
I.8 Doug Roble. Used with permission.
I.9 Ales Leonardis. Used with permission.
I.10 http://en.wikipedia.org/wiki/File:NASA_Mars_Rover.jpg. NASA. Public Domain.
I.11 Waymo Self-driving car. Licensed under CC BY-SA 4.0.
I.12 http://en.wikipedia.org/wiki/File:NASA_Mars_Rover.jpg NASA. Public Domain.
I.13 http://www.sciencephoto.com. Used with permission.
I.15 Terese Winslow. Used with permission.
I.16 https://en.wikipedia.org/wiki/File:Fraser_spiral.svg. Public Domain.
I.17 Edward H. Adelson. Used with permission.

73

Image Credits

I.18 A. Kitaoka. Used with permission.
I.19 http://commons.wikimedia.org/wiki/File:Ames_room_forced_perspective.jpg. Licensed under CC 2.0. Public Domain.
I.21 My Wife and My Mother-In-Law. William Ely Hill, 1888. Public Domain.
I.22 https://upload.wikimedia.org/wikipedia/commons/b/b5/Rubin2.jpg. Public Domain.
I.23 Associated Press World Wide Photos, 1972.
I.24 https://en.wikipedia.org/wiki/File:Kanizsa_triangle.svg. Licensed under CC BY-SA 3.0.
I.25 Greece National Football Team, 2017. Steindy. Licensed under CC BY-SA 3.0.
I.26 http://hamamatsu.magnet.fsu.edu/articles/microlensarray.html
I.27 Lena. Dwight Hooker, 1973.
I.28 John Wright. Used with permission.
I.29 https://www.mathworks.com/help/matlab/ref/rgb2gray.html
I.30 Fredo Durand. Used with permission.
I.31 Edward Adelson. Used with permission.
I.32 Anton Milan. Used with permission.

74

Image Credits

I.33 Marc Levoy. Used with permission.
I.34 Purchased from iStock by Getty Images.
I.35 Purchased from iStock by Getty Images.
I.36 Purchased from iStock by Getty Images.
I.37 Purchased from Shutterstock.com.
I.38 Jian Wang. Used with permission.
I.39 Marc Pollefeys. Used with permission.
I.40 https://vision.middlebury.edu/stereo/data/scenes2001/. Tsukuba Stereo Dataset.
I.41 https://builtin.com/robotics/automotive-cars-manufacturing-assembly

75

Acknowledgement: I thank Nisha Aggarwal and Jenna Everard for proof reading this monograph.