Advanced Topics in Computer Vision and Image Processing

Lecture 01
Introduction to Computer Vision

Military College of Signals

Asim D. Bakhshi
asim.dilawar@mcs.edu.pk

October 30, 2021

Contents

1 The Scope of the Computer Vision

1.1 Puzzle of Human Visual System

We can perceive the vividness of the 3D structure of our environment effortlessly. Our eyes and minds can jointly segment details of a particular object from its background. We can count number of people in a portrait easily and name each one of them and might to emotion recognition from face expressions. However, human visual system still presents lots of challenges to perceptual psychologists which they describe through various optical illusions. You can visit some fun pages with striking illusions here:

- https://michaelbach.de/ot/
- https://www.illusionsindex.org/
- http://www.ritsumei.ac.jp/ akitaoka/index-e.html

1.2 The Ideal of Machine Vision

The ideal of machine vision is to develop mathematical techniques for recovering shape and appearance of the objects from images. There are very reliable techniques trying to approach that ideal these days, however, a lot of it is still a dream to develop a machine vision system that has the visual power of a toddler.

1.3 Computer Vision Being an Inverse Problem

An inverse problem in science is the process of calculating from a set of observations the causal factors that produced them. Computer vision, in this sense, is an inverse problem. We seek to design a solution where a machine can recover some unknowns given insufficient information. Broadly speaking we have following three choices to proceed:

- Modelling the physics or geometry of the visual world.
- Probabilistic models.
- Learning from large set of examples.

However, modeling the visual world is a far complex task than modeling the vocal machinery that produces spoken sounds.

Unlike computer vision, the field of computer graphics work through forward models which try to achieve motion and animation of objects, depict reflection and scattering of light over the surfaces and projections onto a flat or curved image plane.

1.4 Applications of Computer Vision

The applications of computer vision are so many that it is very hard to produce and exhaustive list. They are spread all over from medical imaging to variety of industrial applications such as machine inspection and robotic vision. There are consumer level as well as specialist applications. In fact, computer vision is so pervasive that engineering and technology students are already familiar with them.

2 Brief History of Computer Vision

2.1 The Biological Vision

History of vision can go back to about 543 million years ago. We can speculate that the earth was mostly water and there were a few species of animals floating around in the ocean. Animals didn't move around much and didn't have eyes. Then something really remarkable happened around 540 million years ago. From fossil studies zoologists found out that within a very short period of about ten million years, the number of animal species underwent an explosion. It went from a few of them to hundreds of thousands. It is very hard to speculate the reasons for this increase. There were many theories but for many years, this mystery known as Cambrian explosion was termed evolution's Big Bang by evolutionary biologists.

In 2001, an Australian zoologist Andrew Parker proposed one of the most convincing theory from the studies of fossils he discovered from around 540 million years ago. This was the time when the first animals developed eyes and the onset of vision triggered the Cambrian explosion. Animals could suddenly see and once you can see life becomes much more proactive. Some predators went after prey and prey have to escape from predators so the evolution or onset of vision started an evolutionary arms race. Animals had to evolve quickly in order to survive as a species and that was the beginning of vision in animals. Vision finally developed into the biggest sensory system of almost all animals, especially in intelligent animals such as humans, almost 50% of the neurons in our cortex are involved in visual processing. It is the biggest sensory system that enables us to survive, work, move around, manipulate things, communicate, entertain, and many things.

2.2 Camera Obscura

Camera obscura (from Latin, meaning "darkened room")[1] is a device in a shape of a box or a room that lets the light through a small opening on one side and projects it on the other. In this simple variant, image that is outside of the box is projected upside-down. More complex cameras can use mirrors to project image upwards and right-side up and they can also have lenses. Camera obscura is used as an aid for drawing and entertainment.

Camera obscura is a very old device. Oldest mention of its effect is by Mozi, Chinese philosopher and the founder of Mohism, during the 5th century BC. He noticed that an image from camera obscura is flipped upside down and from left to right as a result of light's moving in straight line. The Greek philosopher Aristotle noticed in 4th century that light from a sun eclipse that passes through holes between the leaves, projects an image of an eclipsed sun on the ground. Passing of light in the straight line also noticed Euclid 4th century BC and Theon of Alexandria in 4th century AD. Anthemius of Tralles, which designed the Hagia Sophia, used a type of camera obscura in his experiments in 6th century. Al-Kindi, Arab philosopher, mathematician, physician, and musician, performed experiments with light and a pinhole in 9th century and proved again behavior of light.

All these scientists experimented with a small hole and light but none of them suggested that a screen is used so an image from one side of a hole in surface could be projected at the screen on the other. First one to do so was

[1] Script for this subsection is reproduced verbatim from http://www.photographyhistoryfacts.com/photography-development-history/camera-obscura-history/ for instructional purposes only, since these notes are not meant to be published as an original work.

Alhazen (also known as Ibn al-Haytham) in 11th century. He was a scientist, mathematician, astronomer and philosopher, he wrote the Book of Optics and, among other things, he invented camera obscura and pinhole camera. At about the same time, Chinese scientist Shen Kuo experimented with a camera obscura. He described it geometrically and even used it explain some effects that were mentioned couple centuries ago but were attributed to the geographic characteristics of the area. As described by Roger Bacon, English philosopher, camera obscura was used in 13th century for safe observation of sun eclipse. Arnaldus de Villa Nova, an alchemist, astrologer and physician, used camera obscura at the same time as a projector for entertainment. Artists started using camera obscura in 15th century. Leonardo da Vinci talks about camera obscura in his "Codex Atlanticus", a twelve-volume bound set of his drawings and writings where he also talked about flying machines, weaponry and musical instruments. Giambattista della Porta, Italian scholar, improved camera obscura by adding it a lens at the place where light enters the box. He also used camera obscura to explain how human eye works. German astronomer Johannes Kepler uses term "camera obscura" for the first time in history in 1604. Johann Zahn, writer of "Oculus Artificialis Teledioptricus Sive Telescopium", writes in his book in 17th century about camera obscura and magic lantern among other optical instruments. In 18th century Conte Francesco Algarotti writes his book "Saggio sopra Pittura" and dedicates a whole chapter to the use of camera obcura (or how he calls it "camera ottica" ("optic chamber")) in painting.

Early models were large and consisted of a literal room or a tent (Johannes Kepler used a tent one.) Later more portable variants were invented. They were wooden boxes that had a lens instead of pinhole which can be moved to provide a focus. They also had a mirror that rotated image and a screen onto which an image was projected. These cameras were basis for early photographic cameras.

2.3 Electrophysiological Developments

In the mean time biologists started studying the mechanism of vision. One of the most influential work in both human vision where animal vision as well as that inspired computer vision is the work done by Hubel and Wiesel in the 50s and 60s using electrophysiology [1].

They chose to study cat brain which is more or less similar to human brain from a visual processing point of view. They placed some electrodes in the back of the cat brain which is where the primary visual cortex area is and tried to record neural activation. They learned that there are many types of cells in the primary visual cortex part of the cat brain but one of the most important collections of cells are the simple ones that respond to oriented edges when they move in certain directions. Of course there are more complex cells as well but by and large they discovered that visual processing starts with simple structure of the visual world, i.e., oriented edges and as information moves along the visual processing pathway the brain builds up the complexity of the visual information until it can recognize the complex visual world.

2.4 The First Computer Vision PhD Thesis: Block World

Block World is a set of work published by Larry Roberts which is widely known as one of the first PhD thesis of computer vision where the visual world was simplified into simple geometric shapes and the goal is to be able to recognize and reconstruct what these shapes are.

2.5 One Summer Expanded to Fifty Years

In 1966 there was a now famous MIT summer project called ”The Summer Vision Project.” The goal of this was to make ”an attempt to use our summer workers effectively in a construction of a significant part of a visual system.”

That was an ambitious goal and fifty years have passed in this pursuit. The field of computer vision has blossomed from one summer project into an army of thousands of researchers worldwide still working on some of the most fundamental problems of vision. We still have not yet solved vision but it has grown into one of the most important and fastest growing areas of artificial intelligence.

2.6 David Marr's Stages of Visual Representation

David Marr, an MIT vision scientist, wrote an influential book in the late 70s about what he thought vision is and how we should go about computer vision and developing algorithms that can enable computers to recognize the visual world [2]. His idea was that in order to take an image and arrive at a final holistic full 3d representation of the visual world we have to go through a chain of processes.

- The first process is what he calls ”primal sketch”. This is where mostly the edges, the bars, the ends, the virtual lines, the curves, the boundaries are represented. This is very much inspired by what neuroscientists had already argued. As you remember, Hubel and Wiesel told us the early stage of visual processing has a lot to do with simple structures like edges.
- Then the next step after the edges and the curves is what David Marr calls ”two-and-a-half D sketch”. This is where we start to piece together the surfaces, the depth information, the layers, or the discontinuities of the visual scene.
- Eventually we put everything together and have a 3d model, hierarchically organized in terms of surface and volumetric primitives and so on.

That was obviously a very idealized thought process of what vision is and this way of thinking actually has dominated computer vision for several decades and is also a very intuitive way for students to enter the field of vision and think about how we can deconstruct the visual information.

2.7 Generalized Cylinder and Pictorial Structure, 1970s

Another very important seminal group of work happened in the 70s where people began to ask the question: "how can we move beyond the simple block world and start recognizing or representing real world objects?"

Two groups of scientists that propose similar ideas: one is called "generalized cylinder" [3] and the other is called "pictorial structure" [4].

The basic idea is that every object is composed of simple geometric primitives; for example a person can be pieced together by generalized cylindrical shapes or a person can be pieced together by critical part in their elastic distance between these parts. Either representation is a way to reduce the complex structure of the object into a collection of simpler shapes and their geometric configuration.

2.8 3D Object Recognition from 2D, 1980s

These work have been influential for quite a few years and then in the 80s David Lowe proposed a computational way to reconstruct or recognize the visual world from simple world structures [5]. He tried to recognize razors by constructing lines and edges and and mostly straight lines and their combination.

So there was a lot of effort in trying to think what what are the tasks in computer vision in the 60s 70s and 80s and frankly it was very hard to solve the problem of object recognition. All these ambitious attempts pretty much remained at the level of toy examples or just a few examples. Not a lot of progress have been made in terms of delivering something that can work in real world.

2.9 From Object Recognition to Segmentation

As people thought about ways to solving problems of vision, one important question arose that if object recognition is too hard, maybe we should first do object segmentation, which is the task of taking an image and grouping the pixels into meaningful areas. We might not know the pixels that group together is called a person, but we can extract out all the pixels that belong to the person from its background and call it image segmentation.

An early seminal work used a graph theory algorithm for the problem of image segmentation [6]. Another problem that made some headway in at the same time was face detection [7].

Around 1999-2000, machine learning techniques, especially statistical machine learning techniques started to gain momentum. These are techniques such as support vector machines, boosting, graphical models, as well as the first wave of neural networks. One particular work that made a lot of contribution was using AdaBoost algorithm to do real-time face detection by Paul Viola and Michael Jones [8]. There's a lot to admire in this work. It was done in 2001 when computer chips were still very very slow but they're able to do face detection in images in near-real-time and after the publication of this paper in five years time, Fujifilm rolled out the first digital camera that has a real-time face detector.

2.10 Feature Based Object Recognition, 21st Century

One of the very influential way of thinking in the late 90s till the first decade of 21st century was feature based object recognition.

2.10.1 Scale Invariant Feature Transform (SIFT)

The seminal work was by David Lowe called SIFT [9]. The idea was to cater for angles, occlusion, viewpoint, lighting and intrinsic variation of the object itself by assuming that there are some parts of the object, i.e., some features, that tend to remain invariant to changes. The task of object recognition began with identifying these critical features on the object and then match the features to a similar object. This is an easier task than pattern matching the entire object.

2.10.2 Spatial Pyramid Matching

Another example is an algorithm called Spatial Pyramid Matching [10]. The idea was that there are features in the images that can give us clues about which type of scene it is, whether it's a landscape or a kitchen or a highway and so on and this particular work takes these features from different parts of the image and in different resolutions and put them together in a feature descriptor and run support vector machine algorithm on top of that.

2.10.3 HoG and Deformable Part Models

A very similar work which has gained momentum in human recognition is to put together features that can compose human bodies in more realistic images and recognize them. One is histogram of gradients [11] and the other is deformable part models [12].

2.11 The Revolution in Quality and Benchmark Datasets

As we moved into the first decade of the 21st century, the quality of the pictures were no longer an issue, the digital cameras were having better and better data to study computer vision and internet was growing. One key outcome in the early 2000s is that the field of computer vision has defined a very important building block problem to solve. In the early 2000s, we began to have benchmark data set that can enable us to measure the progress of object recognition.

2.11.1 PASCAL VOC Dataset

One of the most influential benchmark data set is called PASCAL Visual Object Challenge [13]. It's a data set composed of 20 object classes, such as train, airplane, person, cows, bottles, cats, and so on. The dataset is composed of several thousand to ten thousand images per category. The performance on detecting objects the 20 object in this dataset has steadily increased.

2.11.2 IMAGENET Dataset

Around that time, computer vision researchers also began to ask a harder question: are we ready to recognize every object or most of the objects in the world?

The question was also motivated by an observation that is rooted in machine learning, i.e., most of the machine learning algorithms may they be graphical model, support vector machine, or AdaBoost, are very likely to overfit in the training process. Part of the problem is that visual data is very complex and models tend to have a very high dimension of input and lot of parameters to fit. When training data isn't enough overfitting happens and we cannot generalize very well.

Motivated by this dual consideration of recognizing the world of all the objects and overcoming the the machine learning bottleneck of overfitting, the IMAGENET project was started [14]. The aim was to put together the largest possible dataset of all the pictures we can find, the world of objects, and use that for training as well as for benchmarking. it was a project that took about three years and lots of hard work. It basically began with downloading billions of images from the internet organized by the dictionary called WordNet which is tens of thousands of object classes and then use some clever trick to sort, clean, label each of the images. The end result is a IMAGENET of almost 15 million or 40 million plus images organized in twenty-two thousand categories of objects and scenes and this is the gigantic, probably the biggest dataset produced in the field of AI at that time and it began to push forward the algorithm development of object recognition into another phase.

Especially important is how to benchmark the progress. Starting 2009, the IMAGENET team rolled out an international challenge called IMAGENET Large-Scale Visual Recognition Challenge and for this challenge they put together a more stringent test set of 1.4 million objects across 1,000 object classes and this is to test the image classification recognition results for the computer vision algorithms.

The error rate on IMAGENET detection is steadily decreasing. By 2012 the error rate was so low that it became on par with what humans can do. By the way a human means a single PhD student who spend weeks doing this task as if he were a computer participating in the ImageNet Challenge.

In the first two years our error rate hovered around 25 percent but in 2012 the error rate was dropped more almost 10 percent to 16 percent even though now it's better but that drop was very significant and the winning algorithm of that year is a convolutional neural network model that beat all other algorithms around that time to win the IMAGENET challenge.

3 Focus of this Course: Data Driven Methodologies

The focus of this course is to follow the developments from this point onward where data driven methodologies took over the field of computer vision by sweep. Starting with the generalized image classification problem, we would learn the design and implementation of convolutional neural network models, another popular name of which is deep learning.

3.1 The Image Classification Problem

The generalized setup for the image classification problem is that your algorithm looks at an image and then picks from among some fixed set of categories to classify that image.

This might seem like somewhat of a restrictive or artificial setup, but it's actual quite general. This problem can be applied in many different settings and in this course we are going to talk about some of these machine vision problems, e.g., object detection or image captioning.

3.2 From Image Classification to Complex Problems

Key point to note is that as we move on from problem to problem, the setup, our so-called data-driven pipeline is going to change a little. In object detection for instance, rather than classifying an entire image as a cat or a dog, we want to go in and draw bounding boxes and say that there is a dog here, and a cat here and a car over there in the background. In image captioning on the other hand, given an image, the system now needs to produce a natural language sentence describing the image.

It sounds like a really hard, complicated and difficult problem, but we'll see that many of the tools that we develop in service of image classification will be reused in these other problems as well.

3.3 Convolutional Neural Networks

One of the paradigm that has really driven the progress of the field in recent years has been the adoption of convolutional neural networks or CNNs which are sometimes called ConvNets.

Since 2015, the winner of ImageNet has been a neural network every year. The trend has been that these networks are getting deeper and deeper each year. So AlexNet was a seven or eight layer neural network depending on how exactly you count things. In 2015 we had these much deeper networks: GoogleNet from Google and VGG, the VGG network from Oxford which was about 19 layers at that time. Then in 2015 it got really crazy and this paper came out from Microsoft Research Asia called Residual Networks which were 152 layers at that time. And, since then it turns out you can get a little bit better if you go up to 200, but you run our of memory on your GPUs.

We'll get into all of that later, but the main takeaway here is that convolutional neural networks really had this breakthrough moment in 2012, and since then there's been a lot of effort focused in tuning and tweaking these algorithms to make them perform better and better on this problem of image classification.

3.4 Why Couldn't CNNs Became Popular Early?

It's true that the breakthrough moment for convolutional neural networks was in 2012 when these networks performed very well on the ImageNet Challenge, but they certainly weren't invented in 2012.

These algorithms had actually been around for quite a long time before that. So one of the sort of foundational works in this area of convolutional neural networks was actually in the 90s from Yan LeCun and collaborators who at that time were at Bell Labs [15]. In 1998 they build a convolutional neural network for recognizing digits. They wanted to deploy this and wanted to be able to automatically recognize handwritten checks or addresses for the post office.

The network would take in the pixels of an image and then classify either what digit it was or what letter it was or not. The structure of this network actually look pretty similar to the AlexNet architecture that was used in 2012.

A question you might ask is if these algorithms were around since the '90s, why have they only suddenly become popular in the last 10 years?

There's a couple really key innovations that happened since the '90s.

3.4.1 Computation

One is computation. Thanks to Moore's law, we've gotten faster and faster computers every year. This is kind of a coarse measure, but if you just look at the number of transistors that are on chips, then that has grown by several orders of magnitude between the '90s and today. We've also had this advent of graphics processing units (GPUs) which are super parallelizable and ended up being a perfect tool for really crunching these computationally intensive CNN models. So just by having more compute available, it allowed researchers to explore with larger architectures and larger models, and in some cases, just increasing the model size, but still using these kind of classical approaches and classical algorithms tends to work quite well.

3.4.2 Data

The second key innovation that changed between now and the '90s was data. These algorithms are very hungry for data. You need to feed them a lot of labeled images and labeled pixels for them to eventually work quite well. In the '90s there just wasn't that much labeled data available. It was also very difficult to collect large, varied datasets. But, in the 2010s with datasets like PASCAL and ImageNet, there existed these relatively large, high quality labeled datasets that were, again, orders and orders magnitude bigger than the dataset available in the '90s. These much large datasets allowed us to work with higher capacity models and train these models to actually work quite well on real world problems.

3.5 The Overarching Ideal of Computer Vision and Open Problems

In computer vision we're in the business of trying to build machines that can see like people. And people can actually do a lot of amazing things with their visual systems. When we go around the world, we do a lot more than just drawing boxes around the objects and classifying things as cats or dogs. Our visual system is much more powerful than that.

As we move forward in the field, there's still a ton of open challenges and open problems that we need to address. We need to continue to develop our algorithms to do even better and tackle even more ambitious problems.

Some examples of this are going back to these older ideas in fact. Things like semantic segmentation or perceptual grouping where rather than labeling the entire image, we want to understand for every pixel in the image what is it doing, what does it mean.

3.6 What You Should Expect from the Course

- Understand the deep mechanics of state of the art algorithms.
- How do the architectural decisions influence training and performance of algorithms.
- Implementing algorithms from scratch in Pyhton or Matlab.
- However in most of the cases, people are not building these algorithms from scratch so you'll be introduced to state of the art software tools such as TensorFlow, PyTorch.
- Lastly optimization differences between CPUs and GPUs.

4 Software and Tools

We'll use following software and tools in this course:

- Python 3.x – programming
- TensorFlow 2.1.0 – designing networks
- Numpy – numerical mathematics, linear algebra
- Pandas – manipulating data
- SciPy – array processing

- OpenCV – computer vision
- Matlplotlib – quick plotting
- Seaborn – advanced plotting

5 What you should do this week

- Install Anaconda; learn about any Python IDE; Spyder is the preferred one
- Use pip utility to install required software
- Read these lecture notes
- Read Chapter 1 of Szeliski, Richard. Computer vision: algorithms and applications
- Explore Google Colab

References

[1] David H Hubel and Torsten N Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". *The Journal of physiology* 160, pp. 106–154, 1962.

[2] David Marr and Tomaso Poggio. "A computational theory of human stereo vision". *Proceedings of the Royal Society of London. Series B. Biological Sciences* 204, pp. 301–328, 1979.

[3] Rodney A Brooks and Thomas O Binford. "Geometric modeling in vision for manufacturing". In: *Techniques and Applications of Image Understanding*. Vol. 281. International Society for Optics and Photonics. 1981. Pp. 141–159.

[4] Martin A Fischler and Robert A Elschlager. "The representation and matching of pictorial structures". *IEEE Transactions on computers* 100, pp. 67–92, 1973.

[5] David G Lowe. "Three-dimensional object recognition from single two-dimensional images". *Artificial intelligence* 31, pp. 355–395, 1987.

[6] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation". *IEEE Transactions on pattern analysis and machine intelligence* 22, pp. 888–905, 2000.

[7] Michael Jones and Paul Viola. "Fast multi-view face detection". *Mitsubishi Electric Research Lab TR-20003-96* 3, p. 2, 2003.

[8] Paul Viola and Michael Jones. "Fast and robust classification using asymmetric adaboost and a detector cascade". *Advances in Neural Information Processing System* 14, 2001.

[9] David G Lowe. "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999. Pp. 1150–1157.

[10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE. 2006. Pp. 2169–2178.

[11] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee. 2005. Pp. 886–893.

[12] Pedro F Felzenszwalb et al. "Object detection with discriminatively trained part-based models". *IEEE transactions on pattern analysis and machine intelligence* 32, pp. 1627–1645, 2009.

[13] Mark Everingham et al. "The pascal visual object classes (voc) challenge". *International journal of computer vision* 88, pp. 303–338, 2010.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". *Advances in neural information processing systems* 25, pp. 1097–1105, 2012.

[15] Yann LeCun et al. "Object recognition with gradient-based learning". In: *Shape, contour and grouping in computer vision*. Springer, 1999. Pp. 319–345.