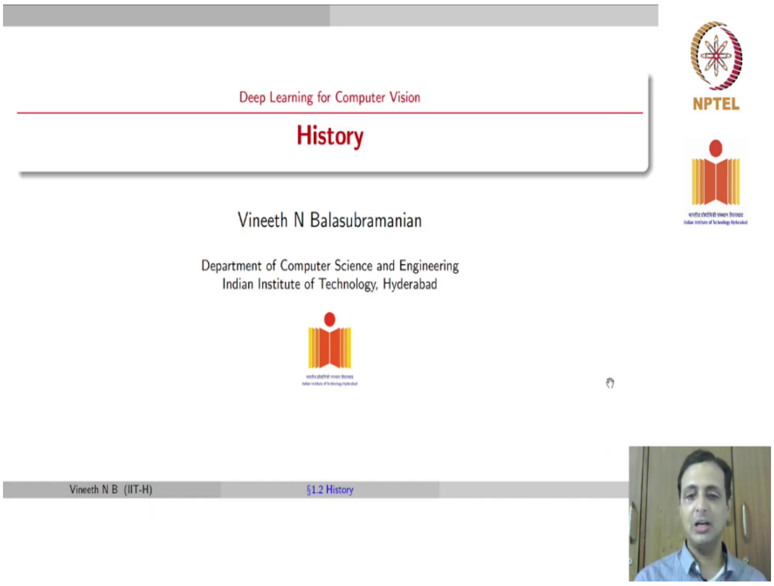
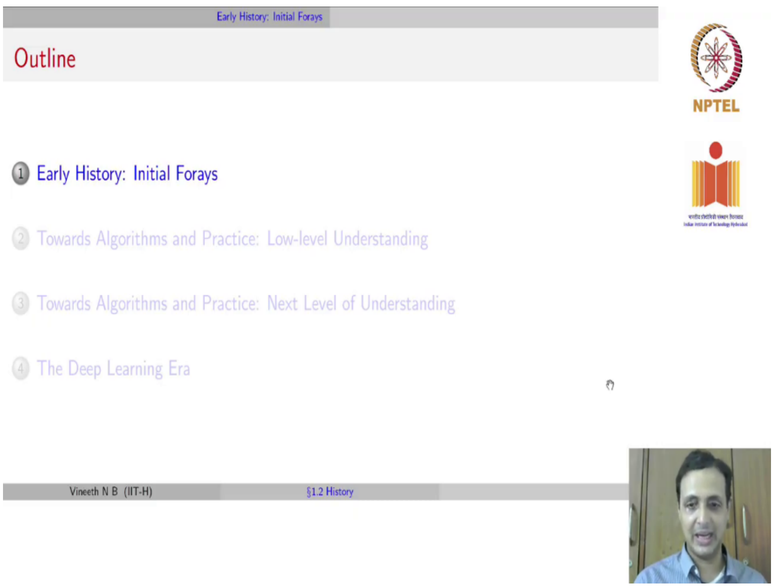
Deep Learning for Computer Vision
Professor. Vineeth N Balasubramanian
Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad
History

(Refer Slide Time: 0:15)



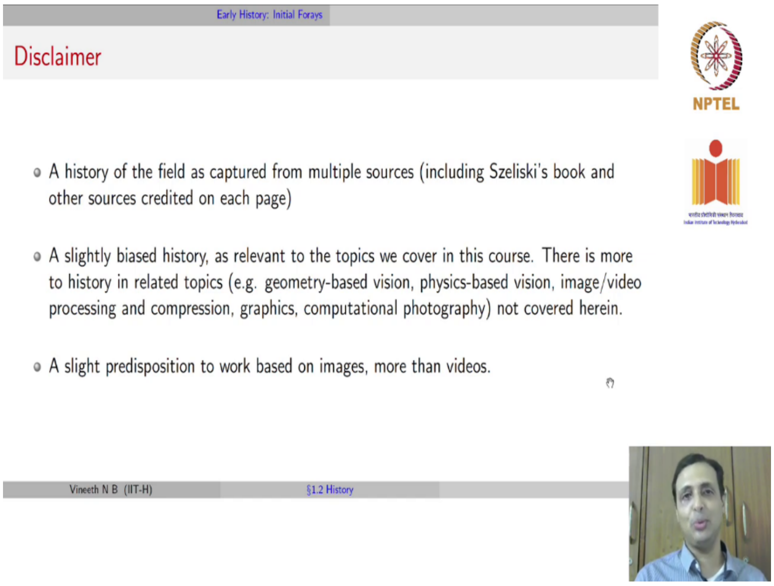
Last lecture we gave an introduction to the course and now we will actually get started with the contents. We will review history of computer vision over the last few decades just to give a perspective of where the field started from and how it is evolved over the last few decades. So, this lecture is structured into four parts.

(Refer Slide Time: 0:35)



We will briefly describe very initial forays in the field in the fifties, sixties and seventies. Then we will talk about affords that contributed to low level understanding of images in the 80s largely, then we will go on to high level understanding the community took up in the 90s and 2000s and of course we will then cover a brief history of deep learning in the last decade or so.

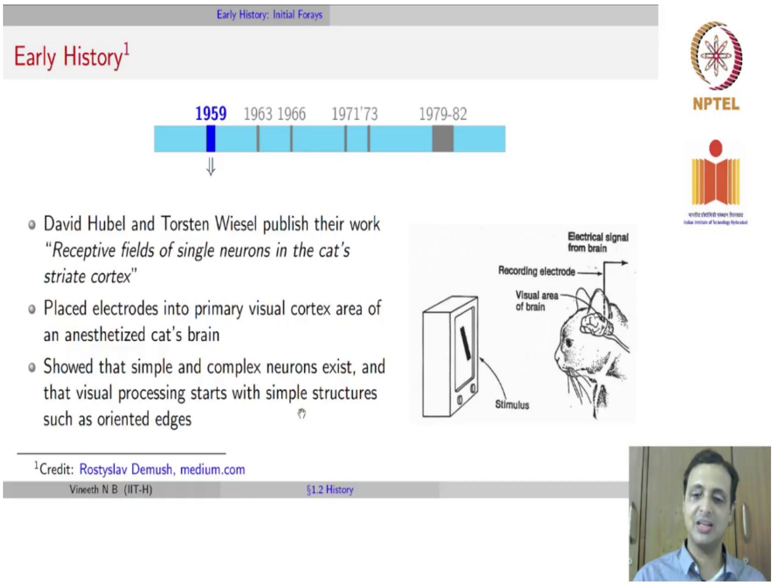
(Refer Slide Time: 1:01)



To start with a disclaimer, this is going to be a history of the field as captured from multiple sources: Szeliski's book as well as many other sources that are mentioned on each of the slides. It may be a slightly biased history from multiple perspectives: 1) perhaps the way I have seen it and I have seen it to be important please bare with that personal bias. 2) It may also be biased to the topics that we cover in the course, may not cover physics-based vision, geometry-based vision in too much detail.

Once again I will refer you to those books that we talked about in the previous lecture if you want to know them in more detail. There is also a slight predisposition to work around images, more that videos but still hopefully this slides gives you a perspective of the field and how it is evolved over the last few decades.

(Refer Slide Time: 1:53)

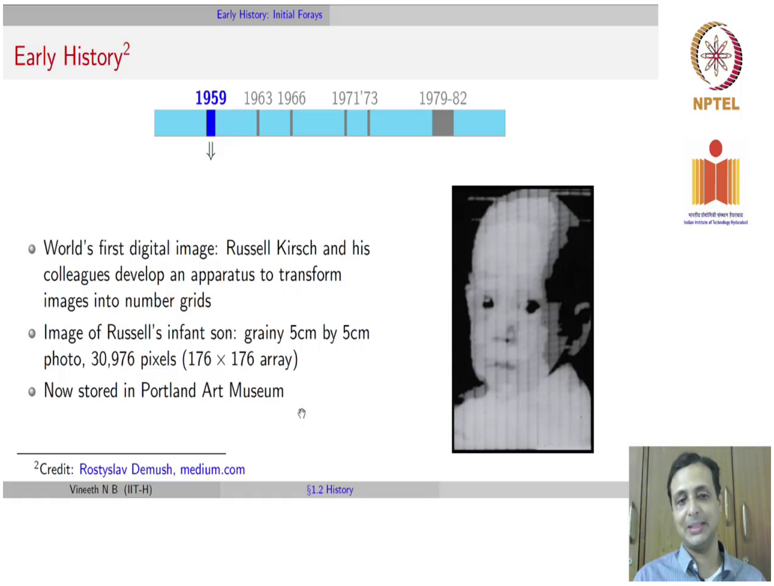


The earliest history of computer vision was way back in the 50s when two researchers David Hubel and Torsten Wiesel published their work called "Receptive fields of single neurons in the cat's striate cortex". So, they conducted multiple experiments to understand how the mammalian vision cortex functions and they took a cat and they did many experiments in this regard but they inserted electrons into a sedated cat and then tried to see how cat's neurons fire with respect to visual stimuli presented to the cat.

Incidentally for quite a long time long time, they could not make headway and accidentally they found that the cat's neuron fire when they switched slides in the projector in front of the cat. They were initially perplexed, but they later realized and that was one of their propositions that the edges created on the screen by the slide that was inserted into the projector was what fired a neuron in the cat.

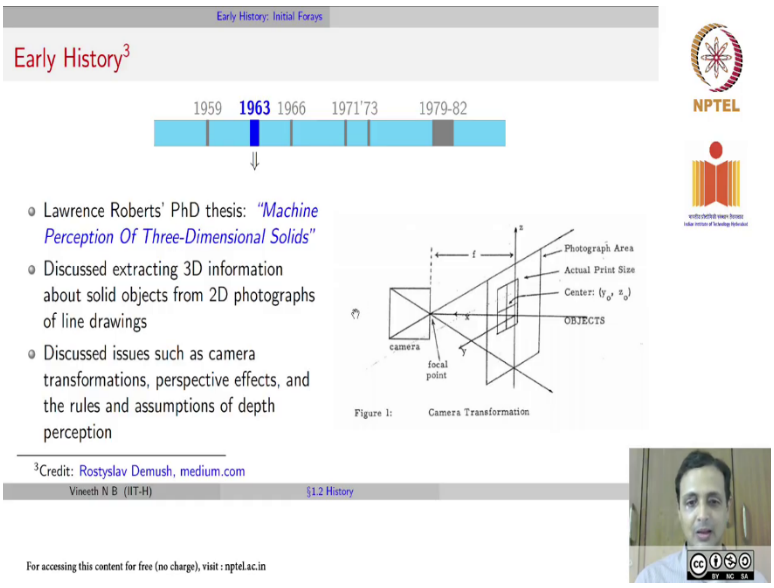
One of the outcomes of their early experiments was that simple and complex neuron exists in the mammalian visual cortex and that visual processing starts with simple structures such as oriented edges. So, Hubel and Wiesel in fact did many more experiments over the next two decades. They actually won the Nobel Prize in 1981 for their work in understanding the mammalian visual cortex. So, this is one of the earliest efforts in computer vision.

(Refer Slide Time: 3:35)



In the same year in 1959, there was actually another major development too, which was by Russell Kirsch and his colleagues were for the first time they represented an image as a set of 1s and 0s. So, representing an image as a number grid is a huge achievement which is something that we inherit to until today and in fact the first image taken was of Russell's infant son which was a 5 centimetre by 5 centimetre photo. About 176 cross 176 array that was captured at that particular time. This is considered as such a big achievement in the field of vision, that this particular photo is still preserved in the Portland Art Museum in the USA.

(Refer Slide Time: 4:24)

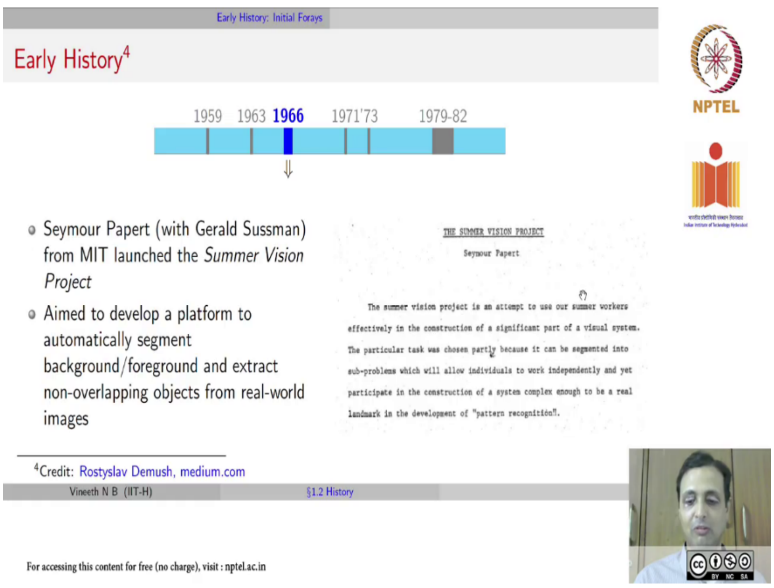


Then in 1963, there was a significant development by a person called Lawrence Roberts and he wrote a PhD thesis on "Machine Perception of 3 Dimensional Solids". The PhD thesis in fact is hyperlinked on this particular slide. So, please take a look at it if you are interested. But I think this thesis had some ideas even beyond its times at that point. So, the thesis discussed by Roberts talked about extracting 3D information about solid objects from 2D photographs of line drawings.

So, if you recall what we spoke in the previous lecture, we said that the aim of computer vision is to understand the 3D world around us from a 2D images that we get or the 2D videos that we get. To some extent this is what was talked about way back in that PhD thesis in the early 60s. So, the thesis discussed issues such as camera transformations, perspective effects, rules and assumptions of depth perception so on and so forth.

Interestingly, Lawrence Roberts moved on from this topic and he is actually more famous for some other big development that all of us owe him for. So, I am going to leave that as a trivial quiz for you to find out. We will talk about that in the next class. But try to find out what Lawrence Roberts is known for and the hint is it is not for anything in computer vision, but it is a huge technological development that all of us today owe him for. Take a look at it and try to find it out before the next lecture.

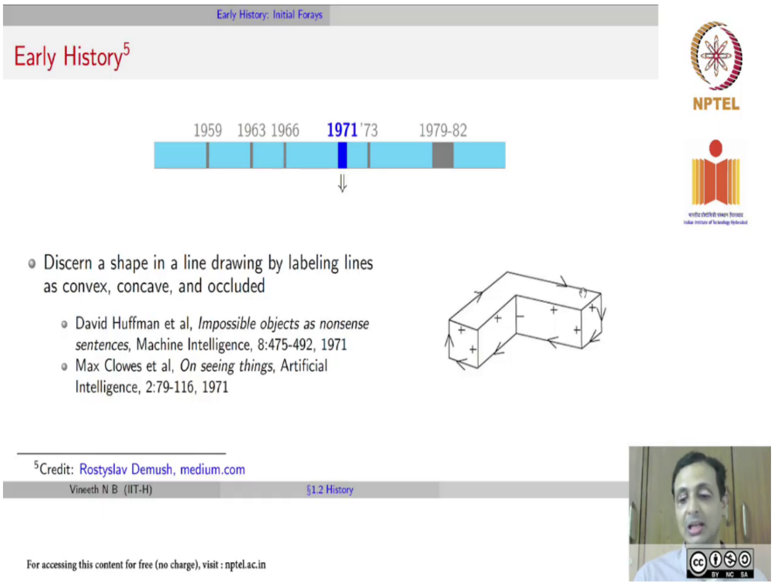
(Refer Slide Time: 6:06)



Subsequently in 1966, one of one of the earliest efforts in trying to come up with systems for computer vision which happened in MIT in 1966 by Papert and Sussman who decided the they could use a bunch of their summer interns to develop an end to end system for computer vision. They thought they could take a few summer interns and develop a platform to automatically segment foreground and background and extract non-overlapping objects from real world images and this is something that they thought they could achieve within a summer.

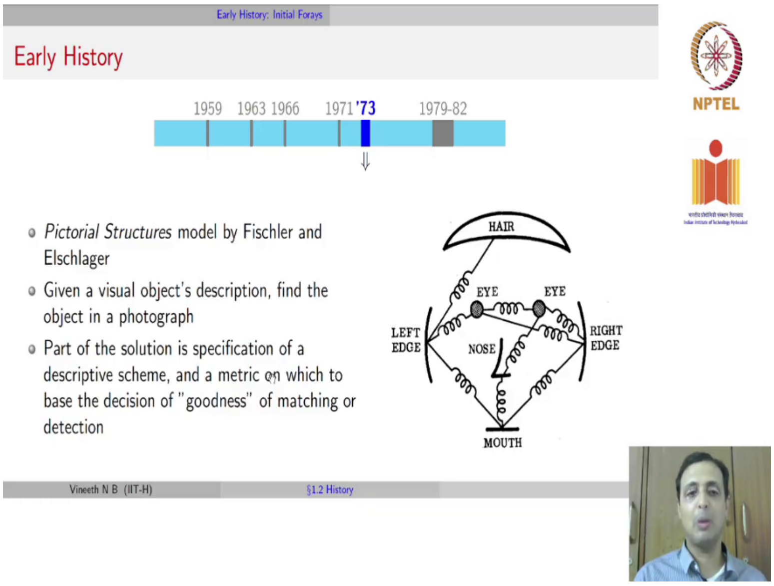
So, this was actually a note that was written by Papert at that time. Obviously, you and I know now that the project did not succeed rather the project opened up researchers to the fact that this was a very deep problem and it was not something that could be solved in 2-3 months and we still know that this problem, certain aspects of it are solved but many other aspects still remain unsolved.

(Refer Slide Time: 7:13)



Then the years went on and early 1970s, there were also works were people tried to study how lines could be labelled in an image as say, convex, concave or occluded or things of those kind. So, that was one of the effort by Huffman and Clowes in the early in the early 70s.

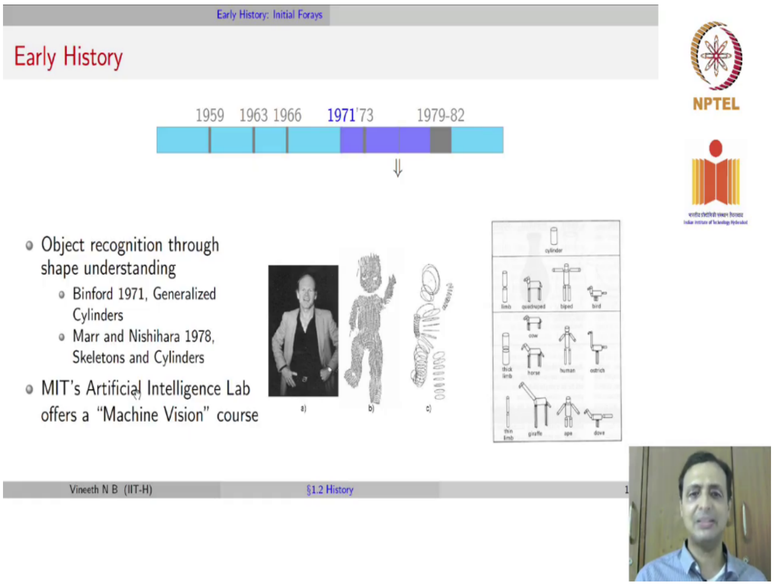
(Refer Slide Time: 7:35)



And in 1973 came an important approach called the Pictorial Structures by Fischer and Elschlager which was again reinvented in the early 2000s, I will talk about that a bit later. But what they talked about there was, they wanted that given a visual object's description that somebody should be able to find out the object in a photograph. So, the part of the solution was to define an object as a combination of individual components and the connections between those components.

And they proposed a solution which firstly a specialization of a descriptive scheme of an object as I said in terms of individual parts and connections between parts. But they also defined a metric on which one could base the decision of goodness of matching or detection based on such descriptive scheme. This is a significant development at this time and a lot of the models that were developed in 2000s inherited this approach to the problem.

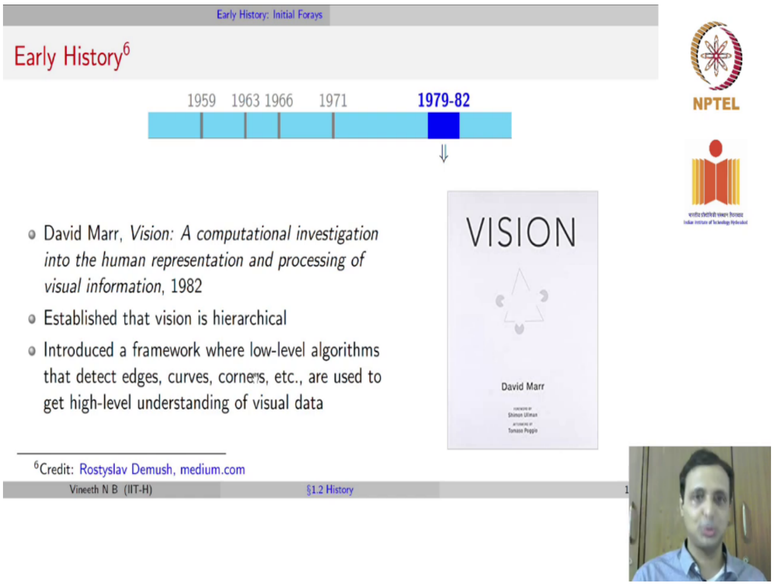
(Refer Slide Time: 8:39)

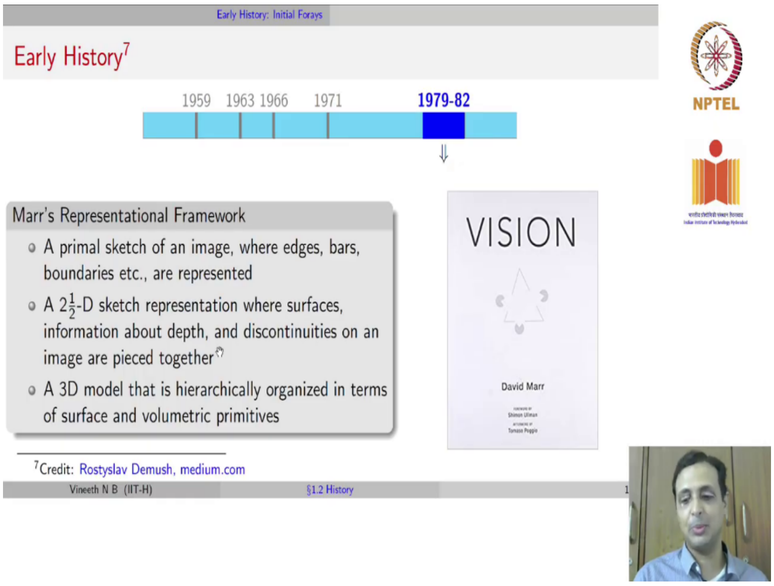


Then between 1971 and 1978, there were lot of efforts that were attempted by researchers and it that period was also known as the "Winter of AI". But at that time many efforts on object recognition using shape understanding, in some sense trying to envision objects as summation of parts. The parts could be cylinders, parts could be different kind of skeletal or skeletal parts was an important effort in that in that time.

So, generalised cylinders, skeletons in cylinders were all efforts at that particular time. And importantly there was also the world first machine vision course offered by the MIT's AI lab in that time in the 1970s. So, I will talk about the applications later, but in the 1970s, also one of the first products of computer vision was developed which was optical character recognition which was developed by Ray Kurzweil who considered a visionary for the field of AI and this was in the 70s again.

(Refer Slide Time: 9:42)





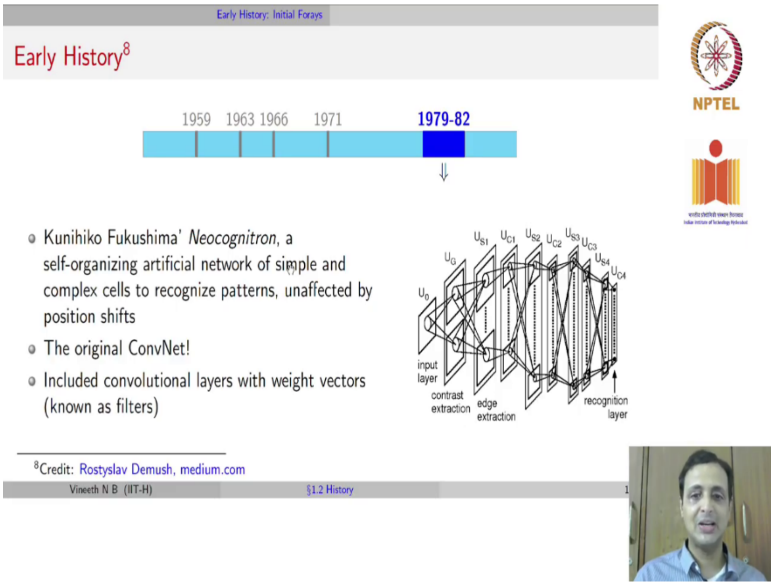
Then between 1979 to 1982 was a again a landmark development for computer vision. David Marr who is research is followed until this, until today. And in fact, the ICCV conference, the International Conference in Computer Vision actually gives out a prize named after David Marr for landmark achievements in computer vision. So, David Marr proposed pretty important framework in his book called 'Vision computational investigation into the human representation and processing of visual information'.

Firstly, he established that vision is hierarchical and he also introduced a framework were low level algorithms that detect edges, curves, corners are then used to feed into a high level understanding of visual data. In particular, his representational framework first had a primal sketch of an image where you have edges, bars, boundaries, etc. Then you have a 2 and a half

D sketch representation where surfaces information about depth, discontinuities are all pieced together.

And finally a 3D model that is hierarchically organized in terms of surface and volumetric primitives. So, to some extend you could say that this also resembles how a human brain perceives information but we will talk about that a bit later. But this was Marr's representational framework which led to a lot of research in subsequent years and decades.

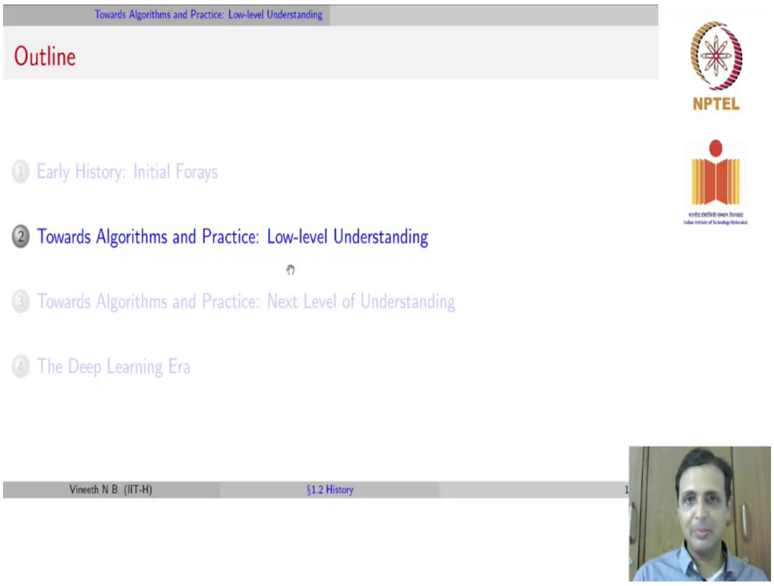
(Refer Slide Time: 11:19)



In the same period around the 80-81 time, there was also a significant development by Kunihiko Fukushima called the Neocognitron which is actually the precursor of convolutional neural networks the day we see today. I think was a significant development for the time and Fukushima introduced a self-organizing artificial network of simple and complex cells to recognize patterns,

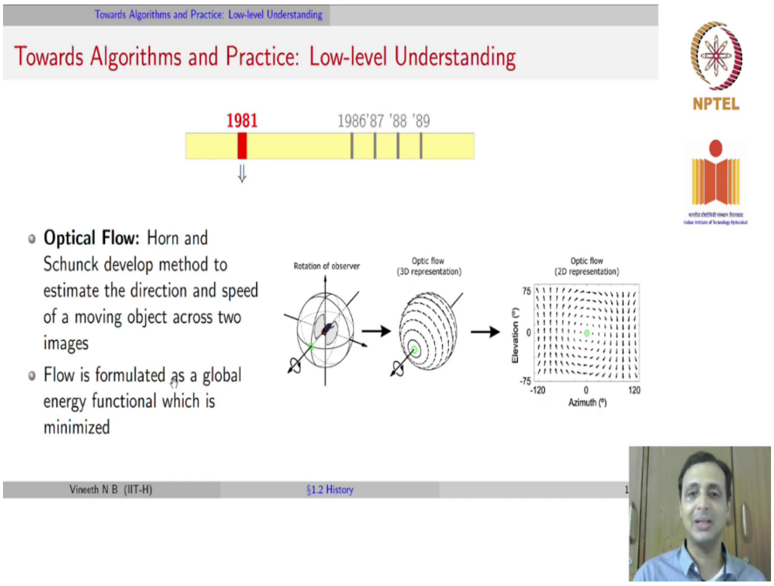
In fact, you can call this the original ConvNet. It also talked about convolutional layers with weight vectors which are also called filters today. That was one of the earliest versions of convolutional neural networks which are used to this day.

(Refer Slide Time: 12:00)



So, that was the initial years and now we will talk about some developments in low level understanding of images which largely happen in the 80s. So we may not cover all of the methods but at least some of the important ones as we go forward.

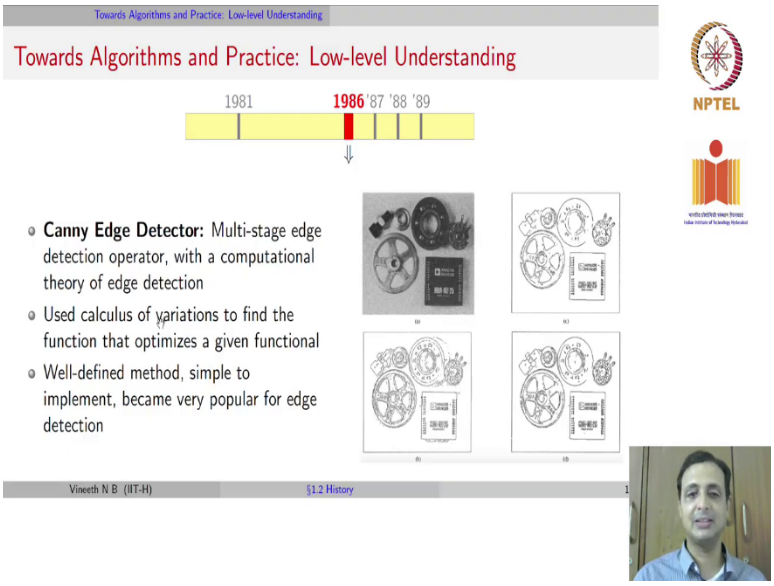
(Refer Slide Time: 12:17)



So, in 1981, there was a very popular method called Optical Flow which was developed by Horn and Schunck and the idea of this method was to understand and estimate the direction and speed is a moving object across two images captured in in a timeline. So, for object moved from position A to position B, then what was the velocity of that object across the two images.

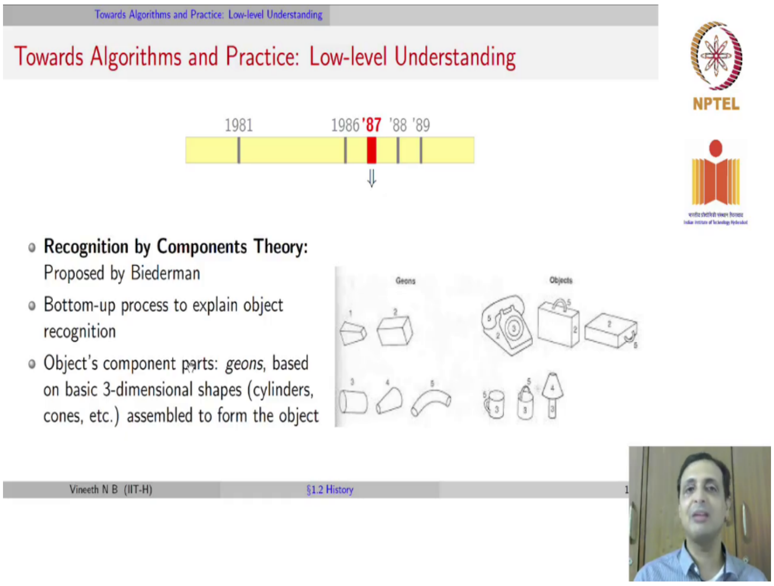
So, flow was formulated as a global energy functional which was minimized and the solution is solution was obtained. And this is the method that was extensively used over many decades especially for video understanding. And I think is still used in certain applications such as say, compression, video compression or other video understanding applications.

(Refer Slide Time: 13:12)



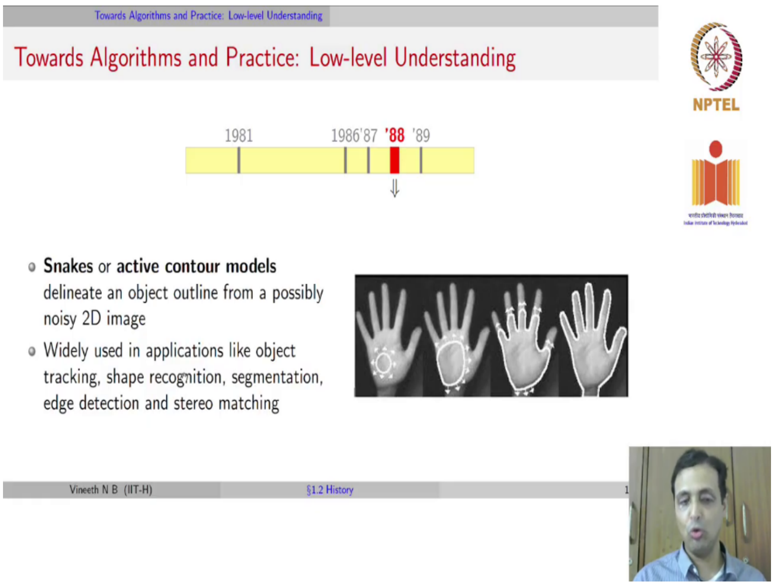
In 1986 came the Canny Edge Detector which was a significant development for Edge Detection. John Canny proposed a multi-staged edge detection operator which is also known as a computational theory of edge detection. It use calculus of a variations to find the function that optimizes a given functional. It was a very well defined principle method, simple to implement and became very very popular for edge detection. So, it was extensively used for many years to detect edges probably until to this day in certain industries.

(Refer Slide Time: 13:47)



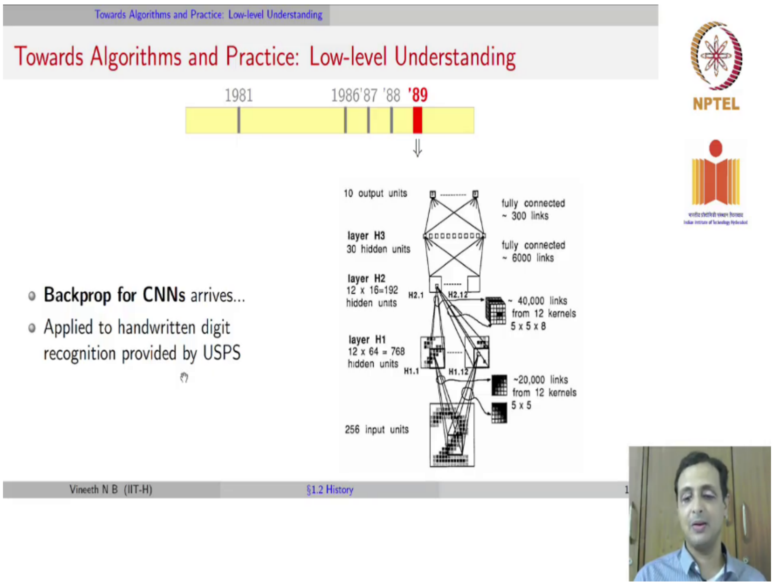
In 1987, there was also the recognition by components theory proposed by Biederman which was a bottom up process to explain object recognition where the object was constituted in terms of parts which were labelled as geons, geons simply meant three dimensional basic three dimensional shapes such as cylinders, cones and so on as you can see in some of this images here which were assembled to form an object. Again this was a theory of visual recognition to see if we could recognise objects in terms of their parts.

(Refer Slide Time: 14:26)



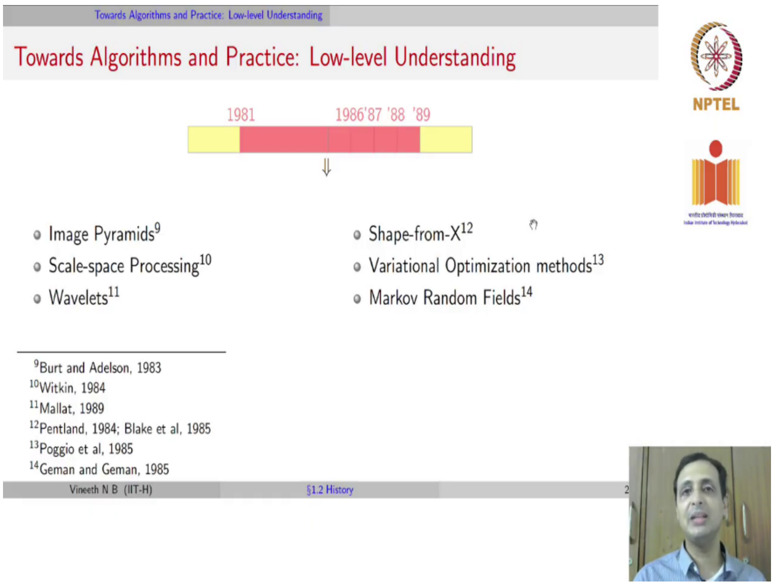
In 1988, came what are known as Snakes or active contour models which helped delineate an object outline from a potentially noisy 2D image. It was widely used in applications like tracking, shape recognition, segmentation, edge detection, so on and so forth.

(Refer Slide Time: 14:48)



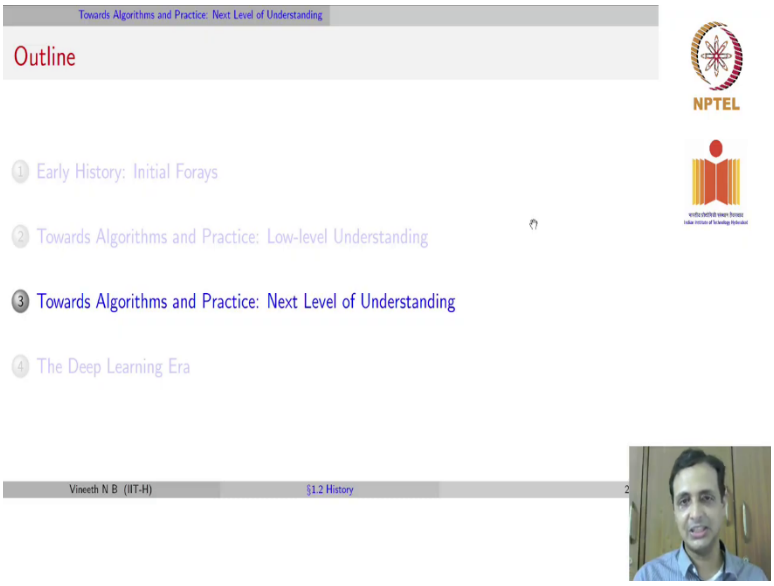
In 1989, was the first version of back propagation for convolutional neural networks. So, it is not necessarily low level visual understanding but I think it happened in the 80s and that is why I am talking about it here and it was applied to hand written digit recognition as we will talk about very soon.

(Refer Slide Time: 15:08)



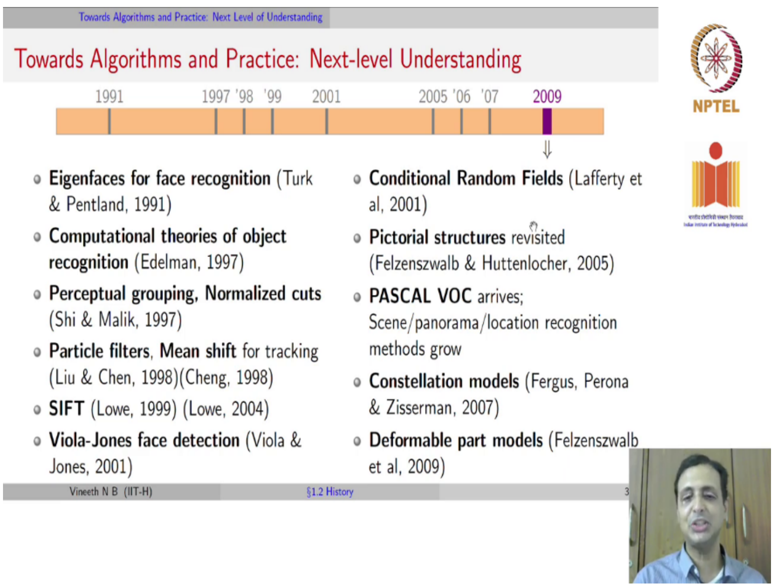
Other things that happened in the 80s where the development of the image pyramids representation of image and multiple scales, scale-space processing, processing of an image at different scales, wavelets which is landmark development at that time. Shape-from-X which is shape from shading, shape from focus, shape from silhouette, basically try to get shape from various aspects of image formation. Variational Optimization methods, Markov Random field, all of these were developed in the 1980s.

(Refer Slide Time: 15:41)



Then came the 1990s where the community stepped into a higher level of understanding beyond low level artefacts such edges or corners or so on and so forth.

(Refer Slide Time: 15:53)



It started with Eigenfaces for face recognition which used a variant of Eigen decomposition for doing face recognition. It happened in 1991 which was successful for face recognition at least in constraints settings. There were also computational theories of object detection by Edelman that was proposed in 1997. Then came Perceptual grouping and Normalized cuts which was a landmark step for image segmentation methods that came in 1997.

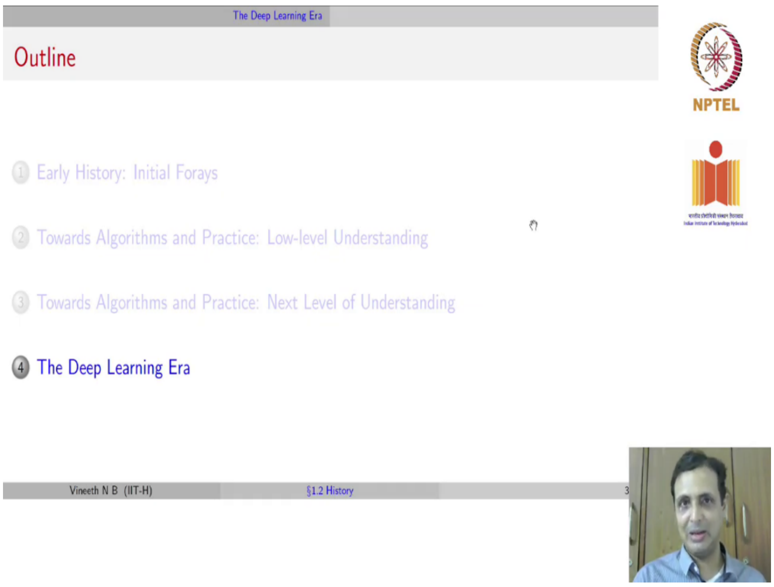
Came Particle filters and Mean shift in 1998, Scale Invariant Feature Transform. We will talk about some of these methods in detail which was an important image key point detector and representation method which was developed in late 90s early 2000s. Then Viola-Jones face detection, again that came in the early 2000s. Conditional Random Fields which was an improvement over Markov Random fields.

Then Pictorial structures, the method proposed in 1973 was revisited in 2005 to develop, they came up with an improved statistical approach to be able to estimate the individual parts and their connections between parts which was called pictorial structures at that time and they actually showed that that could work in practise and give good performance for image matching.

PASCAL VOC which is a data set that is popular to this day actually started in 2005 and around that time between 2005 to 2007, a lot of methods for scene recognition, panorama recognition, location recognition also grew at that time. Constellation models which were part based probabilistic generator models also grew at that time to be able to again recognize objects in terms of parts and how the parts were put together in the whole.

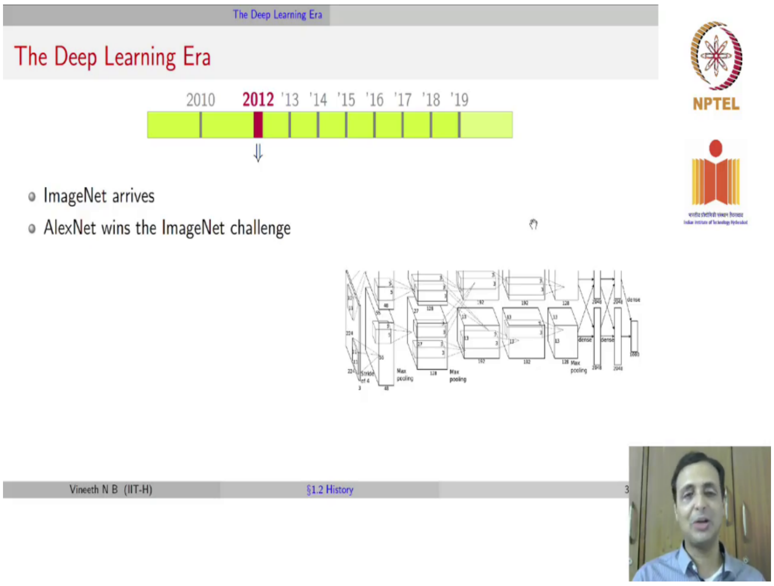
And deformable part models, a very popular approach I think considered one of the major developments of the first decade of 2000 of the twenty first century came in 2009.

(Refer Slide Time: 18:10)



And since then of course, the big developments have been Deep Learning. So, let us briefly review them too.

(Refer Slide Time: 18:17)

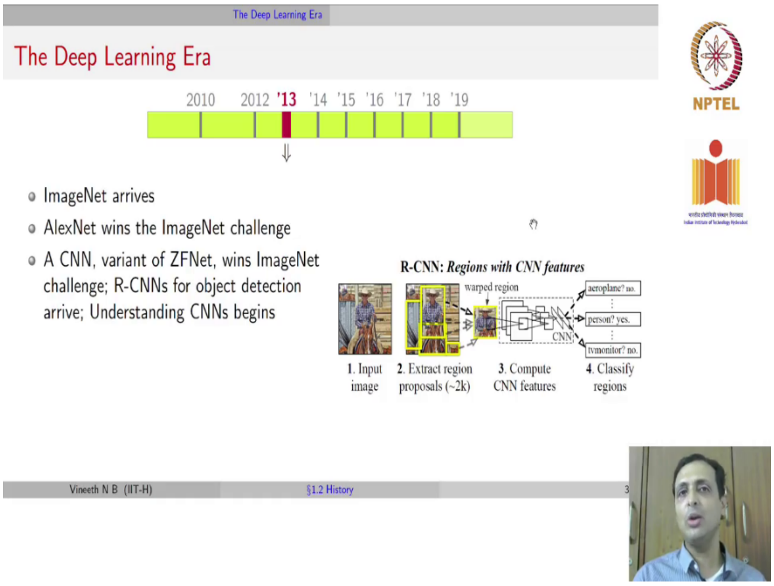


In 2010, the ImageNet data set was developed and the purpose of the dataset was that until then a lot of developments in computer vision relied on lab scale datasets of course, PASCAL VOC dataset changed this to some extent in 2005 and 2006. But many other developments relied on labs scale datasets that were developed in various labs around the world and it did not give a standard way to benchmark methods and compare them across a unified platform, across the unified dataset.

And that is the purpose ImageNet sort to achieve that particular time. So, 2010 was when ImageNet arrived and 2012 was a turning point for deep learning as many of you may be aware, AlexNet won the ImageNet challenge until then all the models that won ImageNet until 2012 were what I mean is shallow models. So, you extracted some features out of the images and then used Machine Learning models such as support vector machines to be able to do object recognition.

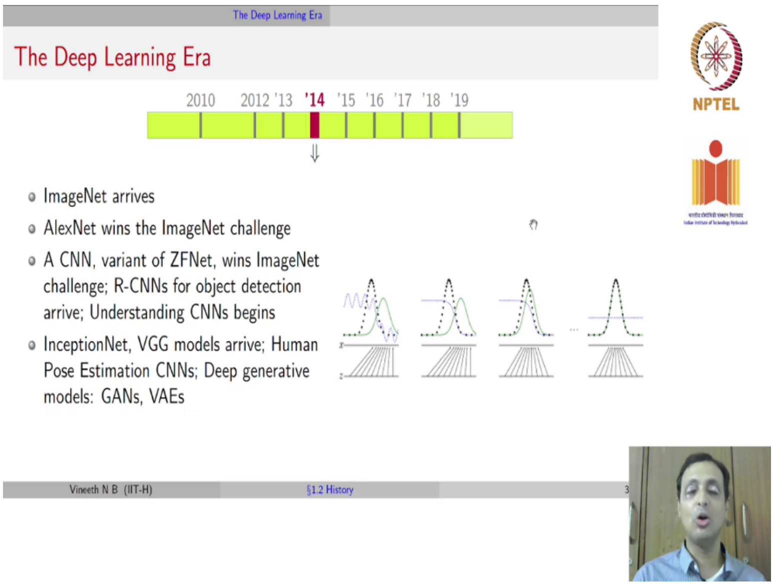
So, in 2012 AlexNet came into the picture and it was the first convolutional neural network that won the ImageNet challenge and it was a significant achievement because it took the accuracy in the ImageNet challenge by a significant amount beyond the previous years best performers. We will talk about the numbers and all of these details when we get to this point in the course.

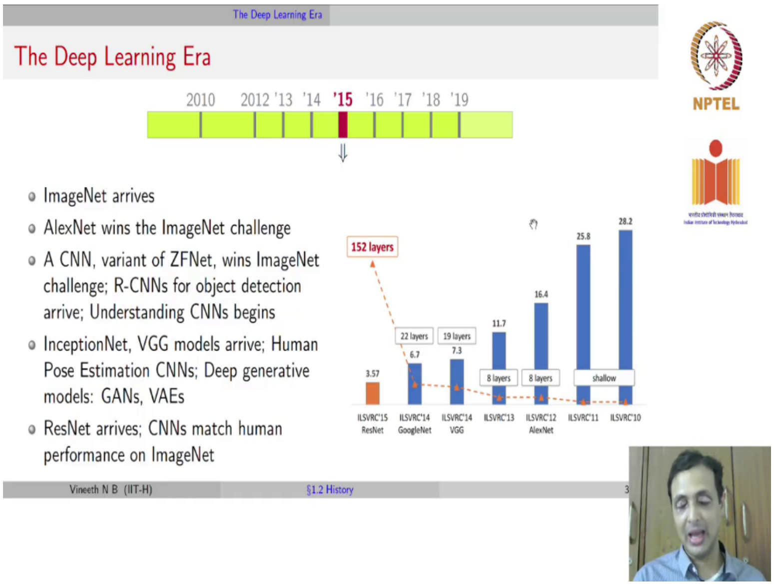
(Refer Slide Time: 19:51)



Then in 2013 came a variant of a convolutional neural network called ZFNet stands for Zeiler and Fergus, it won the ImageNet challenge. Then also regions CNNs or R-CNNs were first developed in 2013 for object detection task and people also started investing efforts in trying to understand how CNNs work.

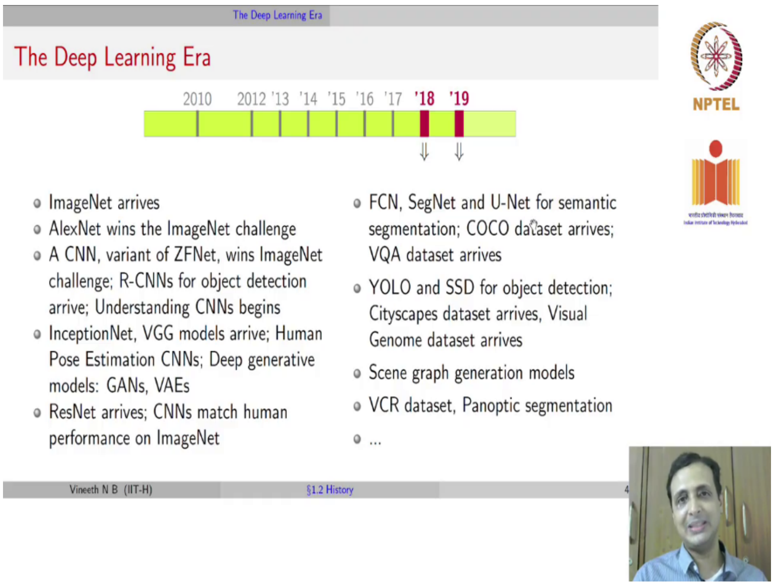
(Refer Slide Time: 20:17)





In 2014, InceptionNet and VGG models arrived. Human pose estimations were developed so, CNN started being used for other tasks beyond just object recognition, deep generative models such as Generative Adversarial Networks GANs and Variational Auto Encoders VAEs also were developed in 2014. In 2015, Residual networks or ResNets arrived and CNNs matched human performance on ImageNet. It was again a landmark achievement.

(Refer Slide Time: 20:53)



2015 also saw segmentation networks that came into the picture. Fully convolutional networks SegNet and U-Net were all developed in 2015 for the task of semantic segmentation or labelling every pixel in an image with a particular class label. The COCO dataset also started appearing at that time and also the first visual question answering dataset VQA dataset was actually developed in 2015.

In 2016, moving beyond region based CNNs for object detection, single stage methods such as You Only Look Once and Single Short Detector, YOLO and SSD were developed. The Cityscapes dataset arrived, the visual genome dataset arrived and 2017 was the start of a higher level of abstraction in understanding images which is scene graph generation. Given an image, how do you understand what is the Scene graph? A person sitting on a horse or a man going on a motor bike, so on and so forth.

And in 2018 and 19, higher levels of abstraction such as the visual common sense reasoning dataset where we try to see if we not only give an answer to a question on an image but can also give a rational to that answer and task such as Panoptic Segmentation have been developed. So, as you can see this journey has focus on going from low level image understanding to higher and higher abstractions of the world we see around us from images.

(Refer Slide Time: 22:34)

From an application stand point, we are not going to walk through every application but at a high level, in the 1970s as I already mentioned, one of the earliest products that was developed was Optical Character Recognition by Kurzweil Technologies by Ray Kurzweil. That was one of the earliest successes of computer vision you can say. In 1980s, most of the industry developments were in machine vision which installed cameras in various manufacturing setups or industrial settings.

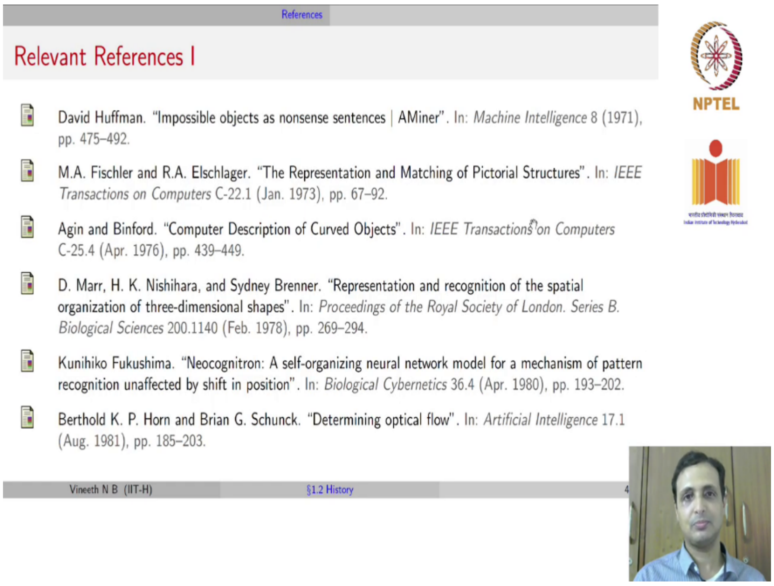
Probably finding defects in processing chips for example or even in smart cameras, where some of these algorithms like edge detection and so on and so forth were embedded as part of the manufacture of cameras itself which I think is known as smart cameras, which I think is a field that is important even today. In 1990s, slowly the applications of vision started growing, machine vision in manufacturing environments continued to grow, biometrics or recognising people from images could be from gait, could be from face, could be from iris, could be from gestures, all of them started growing.

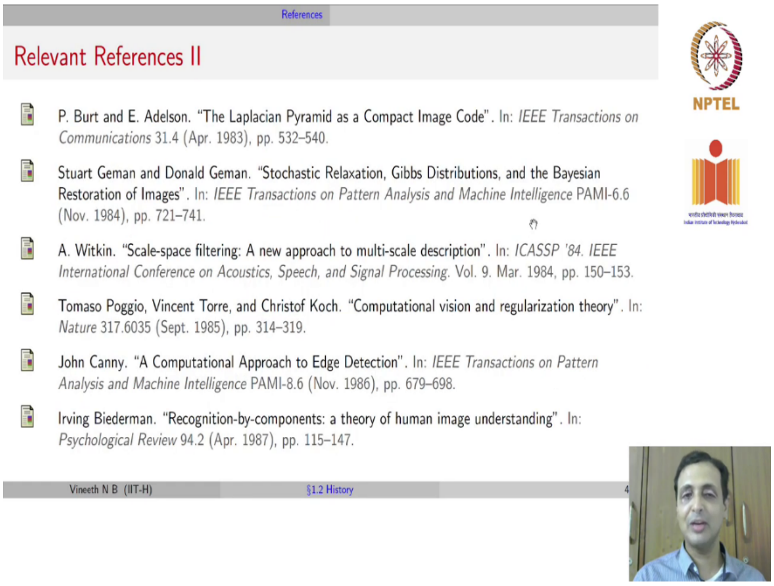
Medical imaging started becoming important. Recording devices, video surveillance, all of them started growing in the 90s. In 2000s, more of all of these, better medical imaging, object and face detection, autonomous navigation started in the mid-2000s, Google Goggles, vision on social media, all of that started in 2000s. And in 2010s, I am not even going to try listing the applications, I think it is grown to a point where vision applications are in various domains all around us.

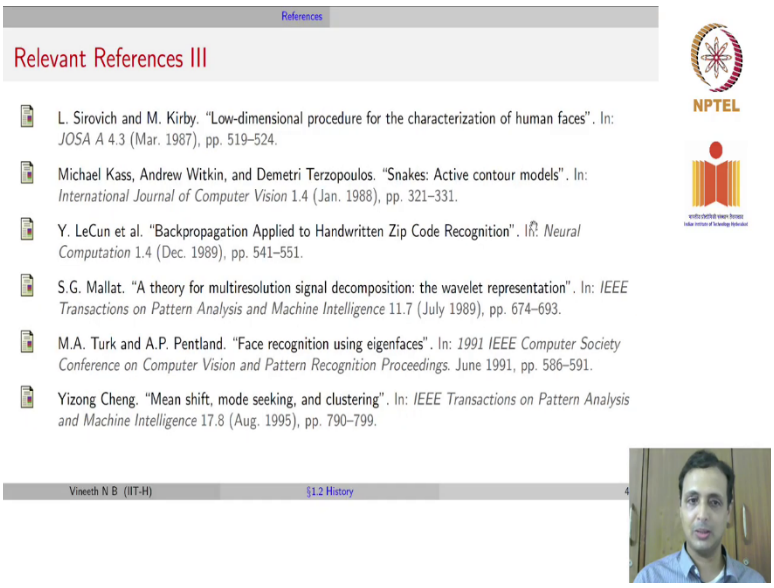
(Refer Slide Time: 24:25)

Hopefully, that gave you a brief perspective of the history of computer vision over the last few decades. I would recommend you to read Szeliskis chapter 1 at this time and also read some of these links that have been shared as part of these slides, every slide had a footnote where the information was taken from. So, go through some of these slides, grow through the links, you will be able to understand how some of these topics grew in specific areas on those links. We will stop here for now and continue with the next topic very soon.

(Refer Slide Time: 25:01)







Here are some references if you like to take a look.