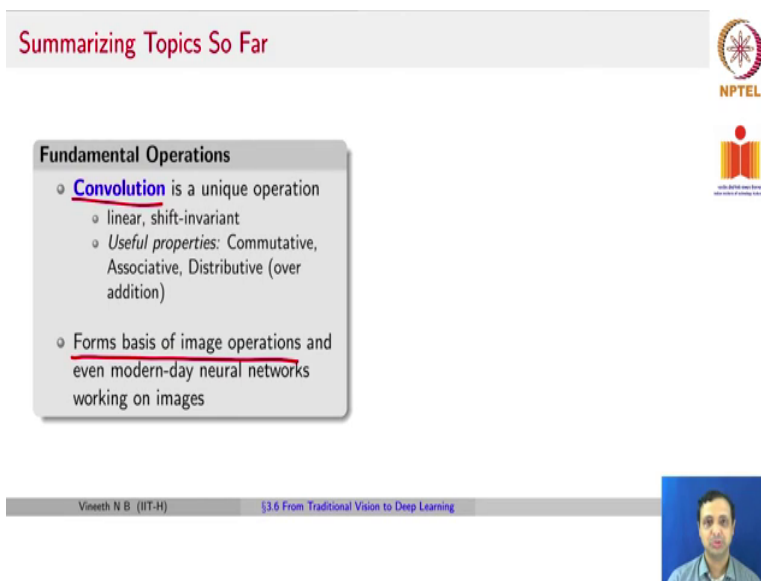


Deep Learning for Computer Vision
Professor Vineeth N Balasubramanian
Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

Lecture 20
Transitioning from Traditional
Vision to Deep Learning

To complete this week's lecture, we will just summarize some of the things that we have seen so far, as we will be transitioning to deep learning from next week. What we have seen so far is a breezy summary of work in computer vision that took two to three decades. So, we have covered a few topics, but we have not covered several more. An important topic that we have probably missed is part based approaches, so on and so forth. Hopefully, we will be able to cover that in a future course, but we will try to summarize the learnings that we have had so far, which will perhaps help us in transitioning to going to deep learning for computer vision.

(Refer Slide Time: 01:04)



The slide is titled "Summarizing Topics So Far" in red text. It features a central box titled "Fundamental Operations" with a list of bullet points. The first bullet point is "Convolution" (underlined in blue), described as a unique operation that is linear, shift-invariant, and has useful properties: commutative, associative, and distributive over addition. The second bullet point is "Forms basis of image operations and even modern-day neural networks working on images" (underlined in red). To the right of the slide are the NPTEL and IIT Hyderabad logos. At the bottom, a video feed shows Professor Vineeth N. Balasubramanian, and a footer bar displays "Vineeth N B. (IIT-H)" and "3.6 From Traditional Vision to Deep Learning".

Summarizing Topics So Far

Fundamental Operations



- Convolution is a unique operation
 - linear, shift-invariant
 - *Useful properties:* Commutative, Associative, Distributive (over addition)
- Forms basis of image operations and even modern-day neural networks working on images

Vineeth N B. (IIT-H) 3.6 From Traditional Vision to Deep Learning

So, one of the things that we learned so far is that convolution is a very unique operation. It is linear, shift invariant. It has useful properties such as commutativity, associativity, it distributes over additions and so on and so forth, so it's very unique in its processing of signals. It forms the basis of image operations and it also forms the basis of neural networks, which are the ones that are used for in computer vision most commonly are known as convolution neural networks. So, convolution still remains used to this day even as part of deep learning.

(Refer Slide Time: 1:48)

Summarizing Topics So Far



Fundamental Operations


- **Convolution** is a unique operation
 - linear, shift-invariant
 - *Useful properties*: Commutative, Associative, Distributive (over addition)
- Forms basis of image operations and even modern-day neural networks working on images

Common Pipeline in Traditional Vision Tasks

- Extract corners or patches in images
→ Extract descriptors
- Use banks of filters, such as Steerable filters or Gabor filters
- Use descriptors for tasks such as retrieval, matching or classification

Vineeth N B (IIT-H)

§3.6 From Traditional Vision to Deep Learning



We have also seen that the common pipeline in traditional vision tasks is given by: we typically extract some points or interest points and images could be edges, or could be key points that have significant change in more than one direction, and we then extract descriptors out of these key points.

This was a common theme if you saw over the last week of lectures at least and we also saw an idea of trying to use banks of filters such as steerable filters or Gabor filters to be able to get multiple responses from a single image and then concatenate them to be able to do any further task or processing. We also saw that these descriptors are useful for tasks such as retrieval, matching or classification.

(Refer Slide Time: 2:48)

Traditional Vision: High-level Abstractions






Image-Level Understanding

- Going from low-level image understanding to aggregation of descriptors
- Banks of filters capture responses at different scales and orientations
- Histograms can be viewed as "encoding" and "pooling"
- Similarities to the human visual system

abstract out the understanding so far it is about each of these we spoke about lower level understanding of descriptors at

Vineeth N B (IIT-H)

§3.6 From Traditional Vision to Deep Learning



If you had to that we had the fact that methods that went from image aggregation

a higher level. So, we use banks of filters to capture responses at different scales and orientations. steerable filters, Gabor filters, so on and so forth. Then there were histograms, which could be considered as doing some form of encoding because you are trying to quantize different key points into a similar scale or even doing some kind of pooling of features to a common cluster centroid or a common codebook element.

So, one could see that there are some similarities here between how this processing was happening to how the processing happens in the human visual system. We at least briefly talked about the various levels of the human visual system, which also bears a similarity of trying to get different kinds of responses at different orientations and scales of the input visual and then trying to assimilate and aggregate them over different levels in the human visual system.

So, there is a similarity here. Although, it was not by design, perhaps it was about solving tasks for computer vision, but there is a similarity about trying to get some lower level features probably features of different kinds with different scales and orientations because choosing only one feature can be limiting for certain applications, so you want to use a bank of different responses and then combine them and be able to assimilate them for further information.

(Refer Slide Time: 04:38)



Traditional Vision: High-level Abstractions


Image-Level Understanding

- Going from low-level image understanding to aggregation of descriptors
- Banks of filters capture responses at different scales and orientations
- Histograms can be viewed as "encoding" and "pooling"
- Similarities to the human visual system

Local Features/Understanding

- Not all spatial regions important, depends on task (stereopsis, motion estimation, instance recognition compared to class recognition)
- Encoding makes features sparse
 - Many words in BoW have zero count
- Operators that detect local features can be viewed as "convolution" followed by some kind of "competition"

Vineeth N B. (IIT-H)
3.6 From Traditional Vision to Deep Learning


Another important thing that we also learned over the last few weeks is that there are applications for which local features are more important. The entire image may not be important, it may be important for certain tasks such as image level matching, maybe an

image level search on one of your search engines or there could be tasks for which only the local features are important, for example, a certain key point or you want to find a correspondence between partially matching images, so on and so forth.

So, it depends on the task, stereopsis is about detecting depth in images, if you want to estimate motion or if you want to recognize an instance of an object, rather than just recognize a class in an image. It depends, as to whether a local region matters or the full image matters.

We also saw that encoding using methods such as bag of words can make your image representation sparse. For example, it is possible that if you had say 10 cluster centers in your k-means for bag of words it is possible that one of your images in your data set may have had only features belonging to three of those cluster centers. The remaining seven cluster centers had no occurrence in that vertical image, which means your image would have a histogram, where for three of those bins, you would have some frequency counts, but the rest of the seven bins will have a 0 count.

That leads to a sparse representation where there are lots of zeros for that particular image. So, encoding can result in that kind of representation for an image. And an important takeaway here is that a lot of operators that detect local features or even global representations of images for that matter can be viewed as performing convolution to get some estimate of features because to detect your key points you need convolution as the key operation that you are relying on, and then that is followed by some kind of a competition.

So, for example, be it the cluster centers, so each of the cluster centers is trying to win votes of different features that correspond to that cluster center, and one of them wins. So, there seems to be some kind of competition or pooling of the result of the convolution operation, which leads to the next step or a higher-level understanding or description of the image.




(Refer Slide Time: 7:13)

Traditional Vision: High-level Abstractions

Representing Images/Regions as Descriptors

- Learn descriptors/representations such that dot product is good enough for matching
- Some invariance to geometric transformations, designed or learned in certain cases

Vineeth N B (IIT-H) §3.6 From Traditional Vision to Deep Learning



So, we also find that the goal so far has been to learn descriptors and representations that make it easy for us to match. You do not want to spend too much time on matching. Of course, we use some intelligence in coming up with matching kernels and so on and so forth, but the key idea is to be able to describe key points, describe images in such a way that the simple dot product or simple matching kernels can be used to be able to match images or parts of images or regions, in images.

These kinds of descriptors have some invariance to geometry, transformations, a certain scale, a certain rotation, certain translation, but in certain cases that are designed in the algorithm in certain other cases, they may have to be learned through other means. This is a brief summary of the topics that we've seen so far put into an abstract manner put into a concise, succinct manner.

(Refer Slide Time: 8:16)

Traditional Vision: High-level Abstractions




Representing Images/Regions as Descriptors

- Learn descriptors/representations such that dot product is good enough for matching
- Some invariance to geometric transformations, designed or learned in certain cases

Moving on to Deep Learning...

Although not by design, Deep Learning seems to build on some of the above principles, but in a learnable manner...we will see soon

Vineth N B (IIT-H) 13.6 From Traditional Vision to Deep Learning



But what we're going to conclude with here is to show that we are going to move to deep learning as I just mentioned, although, not by design, deep learning seems to be building upon some of these principles. Some of these will become clearer when we start discussing these deep learning approaches, but we see that the idea of trying to detect lower level responses of images to different kinds of filters and then aggregating them, and building higher level abstractions and then going to a point of a task where the last representation becomes very simple for a task seems to be very simple, very similar to an idea that deep neural networks also seem to use for solving vision tasks.

Although this may not have been by design, it seems to be similar in the overall structure. But a key difference between all of these methods that we have seen so far and what we are going to see with deep learning over the next remaining weeks of this course is that in deep learning all of this is done in a learnable manner rather than we having to decide, which key point should I use, should I use SIFT or should I use SURF.

Which descriptor should I use? Should I use oriented gradients or should I use GLOH should I use local binary patterns, all of these become design decisions that sometimes become difficult because they may depend on the task, and there is there was no complete knowledge on which kind of a descriptor could be used for which kind of a task.

For example, for face recognition would local binary patterns be always the choice of a feature or could something be used. This kind of a complete understanding of which method to use for which task was not very well known and deep neural networks have in some sense

changed the game there by simulating a similar pipeline, but the entire pipeline is purely learned for a given task at hand. We will see more of this soon, as we go into the next few weeks of lectures on deep neural networks and how they are applied in computer vision.