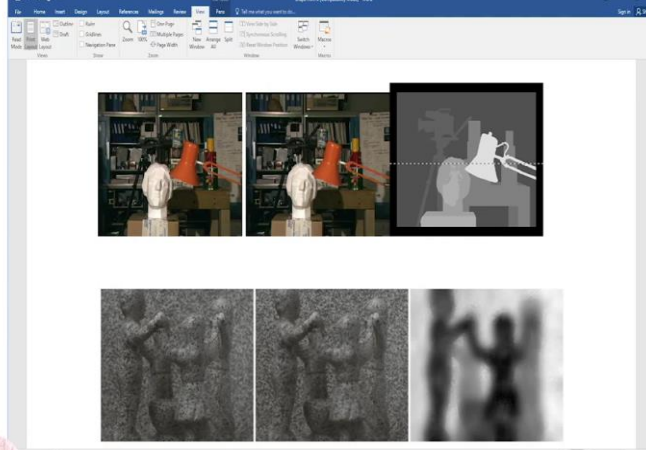



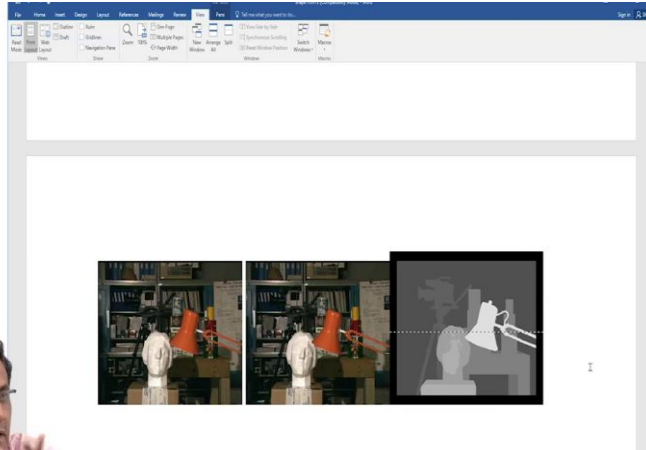

Image Signal Processing
Professor A.N. Rajagopalan
Department of Electrical Engineering
Indian Institute of Technology Madras
Lecture 27
2-View Stereo

(Refer Slide Time: 00:18)



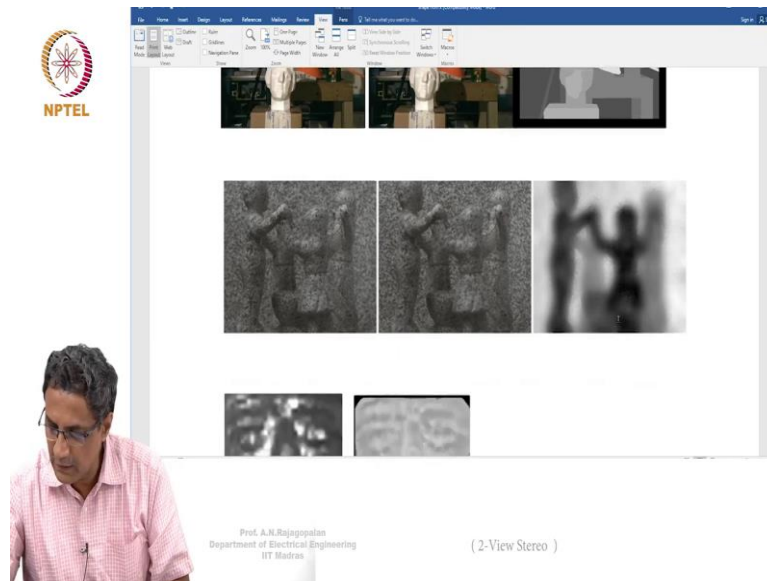
Prof. A.N. Rajagopalan
Department of Electrical Engineering
IIT Madras

(2-View Stereo)



Prof. A.N. Rajagopalan
Department of Electrical Engineering
IIT Madras

(2-View Stereo)



Okay and this is a stereo pair. This is a data set that standard which called a middle bury data set and you know this is supposed to be a stereo pair. So, the stereo pair means that the camera was actually there was a translation. Like I told you only there was a transform cannot be a lower one. Only if there is a translation you can get a parallax effect.

Then as we know right pixel the objects in the front will tend to move more, the objects in the back will tend to move less and therefore that kind of that gives you a cue for dep. Only thing that happens the camera translates and now the right-hand side is really a reconstructed sort of a dep map which looks pretty I mean it may not be that fine to the that level that you want for example, you may say I am not able to see his nose completely. This looks like the whole thing has come out as one thing. So, depends on what application you want.

This is stereo. So, here the camera actually translates. Now if you observe, this one, the earlier one and the earlier one so first one was shape from shading. Next one was photometric stereo. Third one is a stereo. So, in all of these what we are following is really a pinhole model.

A pinhole because none of this was blurred. Did you observe that none of these images were blurred? Only the illumination was changing and some shading was going on but we never had blur in them which means that these all assuming a pinhole model or they are assuming that you are capturing with a lens that has a depth that is large enough so that we do not see any blurring effect at all. Now the next one, this is

something that is called a depth from defocus. So, what this means is that, you take one image with again here the camera is not shifted.

So, the idea is that the each method has its own set of strength and as well as weaknesses. So, stereo the weakness is that, when you kind of move the camera now you have to look for feature correspondence. You have to know where each pixel is gone and that is what gives you cue for depth.

So, that is what makes it hard. Especially read if you have patterns that actually repeat in which case when you do a correlation you could get confused whether this point moved there or it moved right next to it because it could be a repetitive pattern or it could be that you have such a smooth region that hard to tell.

Unless you have a feature, it is very hard to tell where to match it. So, for example if I tell you that from here to here match whom you match. Because they all look the same. So, if you had a features, then it is easy. If you do not have a features that means it is hard to tell where a point is going.

So stereo also has its issue. Another thing is (())(2:56). So, we can that is something that we do not have to talk about now but so similarly read when you come to this one defocusing, depths from defocus. So here is what is done is you take a picture with a lens, with a real aperture camera and assume that there is the depth of field which is in action and this and that you can get varying levels of blur in the image. And you know that suppose I ask you, I give you one image which has a space variant blur.

Let us say that this class, suppose I take a camera, I focus with respect to the front row. It is in my hands what I focus with respect to what kind of looking distance. I focus with respect to the front row. All of you are blurred or let us say I focus with respect to middle row. The guys in the back are blurred. The guys in the front are blurred.

Can you use that one image and tell depth? Or can you devise an algorithm? As a human being you will do all of that but let us say what would you do I mean if you had to do it? Can you devise a method that will use this single image and be able to tell?

It is hard no? Why is it hard? Because of the fact that see you may think that if I had some way to find out if something is more sharp as compared to something else that you could still go wrong. Because of the fact that see it can happen in an image, you find that at some region looks like it is actually blurred but then that could mean you will have to evaluate some kind of a sharpness. You should have something like a sharpness this one.

You should have some operator that will tell you the degree of sharpness. Let us say I have some sort of a tool like that which if I apply on this image right it will actually tell me what level of sharpness is there at every location. But if at some location if I say if it this operator tells me that this has more sharpness as compared to something else which has less sharpness, it does not automatically mean that the one that is more sharp that is actually in focus.

It could very well be that this guy is actually more blurred than the other and all the sharpness is coming out because of the fact that the intensity is below are much more active here as compared to what is happening there. You see that right?

So, I could end up interpreting things wrongly just by having one single image unless you know that it is all uniformly textured. The whole scene has a same texture everywhere then it is okay. Then because you ample like uncertainty about what the image might contain is gone. Because right they are all affected by the same image intensities. Then it is okay.

But that is not decays normally. You have intensities varying all over like me, then this background. So, you cannot, you have to worry about what problem is until like intensity it was like and that could affect what you are focus measure is telling and therefore with a single image it is hard to tell but then if I have 2 or more, then it looks like relative notion comes in now. Because you know in depth in defocus what happens is you take one image and then you change the lens parameters. You stay right there.

You do not move the camera. You take one more and now the idea is that you know some place would have become less blurred, more blurred and that would now start to act as a cube. Because now relative because now scene is out of the way. Because you have that both you know are affected by the same until seeing no? Because you have

not moved the camera and now you know that there is a change in blur probably that could act as a cue.

Because until I seeing this there is no longer influence your idea about what is a degree of blur. That is what a depth from defocus. This in fact write them in things like this I recently read there is a Panasonic which one Lumics GH5S or something that you know 2019 they actually they used this principle in order to do this hunting. So hunting is when you want to do an autofocus no. So, autofocus when a camera does, one of the things by the way it uses a sharpness measure.

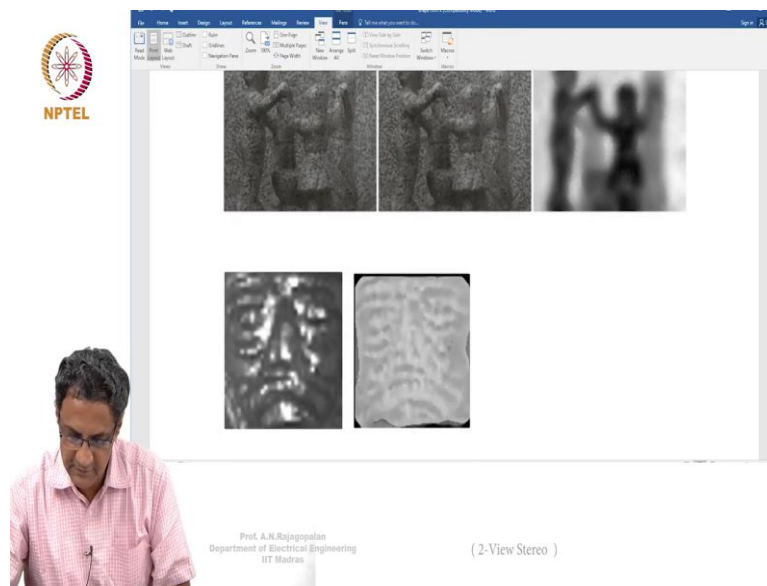
It will capture a bunch of frames quickly. You do not see any of that. We just see the final picture that looks nice. But then it captures very quickly a bunch of frames and by actually by kind of having sort of moving the lens. Does not worry about scaling and all that because all that it needs to know where is things, where are things coming into focus especially with respect to the central region runs of focus measure as required, the sharpness measure quickly but this Panasonic thing, this is what all other phones do.

They capture something like 30 40 frames and then quickly run a measure because relatively you know where things are coming into focus and then you get a freeze those lens settings. But this guy read apparently has run up a 100 xp on that because he just uses 2 images and then right computes, gets an idea roughly about where the object is and then does a bunch of rims around that.

So now he does not capture 50 and all. Does capture probably 10 frames around a place that he senses there is, that is where the object is and then does something very fast on those few frames. So, this is something that I just read somewhere.

Just trying to see where these people uses depth from defocus and all that. So, hunting is one thing read where they actually use for autofocus that but there you know really they are not interested in filing the depth map and all that. This is interested in knowing how to use blur as a cue to know where you want a focus.

(Refer Slide Time: 08:15)



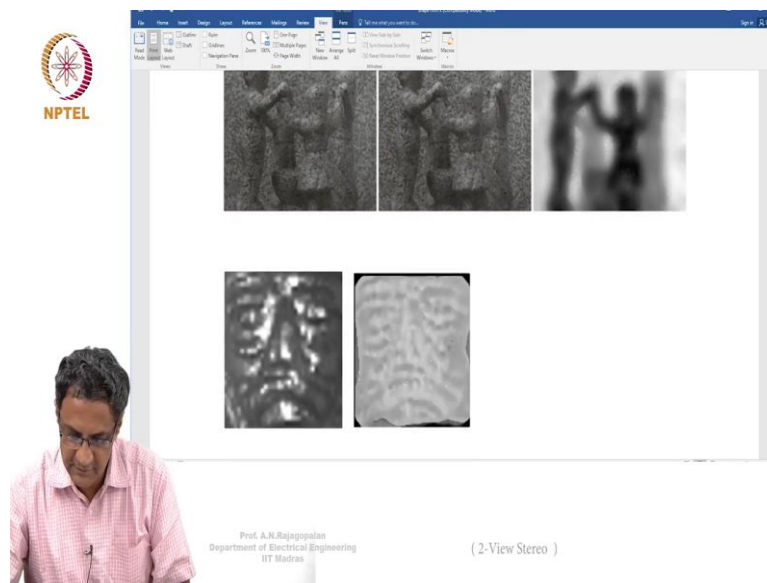
Now, depth from defocus is this and the further going down. This is shape from focus. It is a ring. It is a ring and what is being done is there is this microscope and that microscope it has a stage to keep the ring on the stage and then you move the stage up and down. As you move the stage up and down, so similar to the what we have already said it. So basically, different layers will come into focus right. So, as you can see, not a single frame will be completely in focus.

See here the face is blurred. The background was sharp initially. Then later the face became sharp. The background became blurred. So if I play it again just watch. So basically nowhere, there was no single frame where it could have said that the whole thing is in focus and this you are talking about a ring by the way. So, you can imagine

the depth variation hardly what a few millimetres. That is all you can engrave on a ring. But the blur showing up. It is saying that the object that you have is not flat.

That means somebody actually engraved it. Because if had drawn with a hand, it would have all come into focus but because there is an engraving, there is a depth variation and the depth variation is creating a sense of different degrees of blur as you walk through. Now, this face is what I showed there. Okay so this portion was cropped and the idea is that can you get a views information like this to be able to say what is the object shape? You know let me go back there.

(Refer Slide Time: 09:55)



So, now maybe you can relate. Now you have to relate. This is that face and this is that depth map. This depth map may not be so great to appreciate but I have others. But the catch is this right, think of it, I mean if you cannot do stereo on this, first of all such a shiny surface, you may not even find out where every point is going so hard and then you need a depth map that should be tense because just give you a few points, you cannot even make out that it is a face and thirdly you may not even be able to do stereo because you know you need a baseline of the object that is so small.

There are stereo microscopes but the idea is that can you do using something else? And after all a microscope comes with a depth of field that is only terms of microns. It comes with that. It is not like you have to make an extra effort. It comes with kind of a depth of field and therefore you may want to say this lens formation I know how the

images are formed? Now, can I read use that as a cue to now extract my like I say depth out?

(Refer Slide Time: 10:59)

The slide contains the following handwritten text:

- Active: Kinect, LiDAR
- Passive: Shape from X (Computer Vision)
- co-operative methods
 - pin-hole model
 - Real open camera
- Stereo, defocus, shape from focus, shape from shading, photometric stereo, texture

NPTEL logo is visible in the top left corner. A photo of Prof. A.N. Rajagopalan is in the bottom left. The text '(2-View Stereo)' is in the bottom right.

Now let me just quickly go through just one sort of a derivation for stereo. Before I, so I will start with a pinhole model. What is the typical idea that one uses in a stereo? Let us kind of look at stereo. How does stereo work?

(Refer Slide Time: 11:15)

NPTEL

Stereo, defocus, Maps from focus, Maps from shading, Photometric stereo, texture.

Stereo (pin-hole)

$f(x, z)$

x_c

$(0, 0)$

b

f

z

$\frac{x_c}{f} = \frac{x}{z}$ $x_L = \frac{f \cdot x}{z}$ $\frac{x_n}{f} = \frac{x+b}{z}$

Prof. A.N.Rajagopalan
Department of Electrical Engineering
IIT Madras

(2-View Stereo)

NPTEL

Stereo, defocus, Maps from focus, Maps from shading, Photometric stereo, texture.

Stereo (pin-hole)

$f(x, z)$

x_c

$(0, 0)$

b

f

z

$\frac{x_c}{f} = \frac{x}{z}$ $x_L = \frac{f \cdot x}{z}$ $\frac{x_n}{f} = \frac{x+b}{z}$ $x_n = \frac{f \cdot x}{z} + \frac{f \cdot b}{z}$

$\frac{x_n - x_L}{\text{Output}} = \frac{f \cdot b}{z}$ $z = \frac{f \cdot b}{x_n - x_L}$

Prof. A.N.Rajagopalan
Department of Electrical Engineering
IIT Madras

(2-View Stereo)

So why is it that if you do a correspondence matching you get a sense for depth? So, the way to understand this to sort of imagine that, let us say that you have a point in the scene. Let us say I have a lens. No not a lens. I really have a pinhole okay? And this ray, it travels through the aperture and then hits my lens somewhere here. And this is my focal length, f. This is really a pinhole okay? Stereo is typically for pinhole.

You can also do stereo if it is a real aperture. Nothing stops you. But normal stereo algorithms are all assuming the pinhole. Okay so here is a point p and let me say that and this of course optical axis. This is the aperture.

This is pinhole okay. Now, if you say that right I mean I know this point here and then I want to find out where is p . I can keep traversing on this ray but then I do not know where to stop because the point could have been here and I could have been still guard an intensity there. The point could have been here and I still could have guard an intensity there. So, any point in this ray is equally valid. I do not know where to go back and stop. So in stereo what is normally done is you translate the camera. So, you are here. Here is your pinhole. Now you translate.

Let us say you come down like that and then you take one more. Let me draw one more optical axis here and then you take a second image which basically means that, earlier you had a point, so the image of the point in the first image appeared here. Now, you kind of say translated and this called the base line shift. Okay so the way it stereo typically works is, you do not bring it forward and all. It simply you just translate in plane.

You go like x or you go like y . Typically it is x . You do not do x and y and all. You do not gain anything great by that. So, you will simply do let us say an x shift. So, that is what I mean by shifting it down. Just allow this x direction. So, now if I were to tell a few things, let us say from here to here is your depth and that let us say z and let us say the initial origin let us assume it is here with respect to the first camera centre let us say that is may origin and let us say this p has coordinates, let us say x . so x would be this.

So, this is x . So, x and z with respect to this camera centre and then I have moved by an amount let us say d . That is a baseline shift. I have moved this camera by d . I have come down here and then I see this point here. Now, suppose I assume that there is a shift or something that tells me exactly that this point is gone here.

Now, the same point I can also draw here, instead of looking at this inverted thing I can say that this is my x_L which is my image coordinate in the first. It is called a left right pair typically. Left, right okay? So that is why read x_L and x_R . So, left right.

So, this point this is x_L , I am just marking it on the other side and similarly this is your x_R . So, from here to here which means that I had to go exactly 1 focal length away here somewhere and let us say I have something like x_R . So, from here to here

exactly f alright? Because I am just lurking it on the other side. So here to here this is also f .

Now, if I simply use similarity of triangles right, so what I see is with respect to the first one, I see that x_L by f is equal to, so where is this x_L , so I am looking at this angle right. So that is the same as, this is your origin so x by z or in other words x_L is fx by z . This is something that we know, this is a perspective projection. You know this.

Now, look at this other one. Look at this angle. Now, what I have is x_R by f again because from here to here is exactly f , so x_R by f is equal to now from here to here is actually x plus b , because I have translated further go down by b . So, it is like x plus b by a , z is from here to here. I wrote it little wrongly. z is from here to here okay.

I think I wrote it from there to till the image plane. Not from the image plane you know that is from the aperture. z is there. So, you have x plus b by z . Therefore, x_R is fx by z plus fb by z . Therefore, if you compute a disparity which is x_R minus x_L , disparity is the shift. So, this is called disparity. So, which means that this is the amount by which this pixel has shifted when you shifted the camera. On the image plane how much has it shifted? It shifted by x_R minus x_L and that as you can see is simply fb by z .

Or in another words I can compute z to be fb by x_R minus x_L . So, it is like saying that now that I have a second ray right, if I do a triangulation earlier I went on these ray I did not know when to stop. Now, I have one more point and that is the same pixel right. That is the same point that is appearing here. So, when I go back where these 2 rays intersect is where that p is.

In effect triangulation is what you are doing. But in terms of just simple equation it turns out that if you know f precisely, if you know the baseline shift that means you are actually know by how much you shifted and if you can compute that disparity which will have to come through shift or something which is again hopefully robust then you know that you can actually compute z which basically means that if you can do this for every point in the image for every point you can find where it is gone. Compute this disparity. Then you know you can compute the whole z . But then it is not that easy because you know there will be errors and so on.

Because the correspondence could go wrong or image may not have say this one texture in some places. Things will go wrong in little bit here and there but then overall right people know that okay there is this information which can be tapped. So this is how stereo works but here you actually move the camera. We will look at the other one when we meet on whatever in next class.