

**Image Signal Processing**  
**Professor A. N. Rajagopalan**  
**Indian Institute of Technology, Madras**  
**Department of Electrical Engineering**  
**Lecture 28**  
**Introduction to Shape from Focus**

Last time we saw I just finished to telling you about stereo. So, stereo actually gives you a cue for depth, if you are interested in depth. So, we saw that if you actually translate the camera then you saw that equation right, last time when they derived.

(Refer Slide Time 0:48)

Stereo: Disparity map

$$z = \frac{f \cdot b}{x_r - x_l}$$

Disparity

Focused

Defocusing (optical) / Motion blur

↳ Due to lens.

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

( Introduction to Shape from Focus )

So, what was actually a disparity, so when you talk about stereo, talk about a disparity map as it is called. And we are not going to talk about it in detail, but, so disparity map so we will get this. So, what was that? So, you had an equation that  $z$  that is just dependent on the focal length, the baseline  $X R$  minus  $X L$ , this is called a disparity.

Or in other words, it shows that  $X R$  minus  $X L$  is inversely proportional to  $z$ , which means that points that are closer will have a disparity that is larger and points that are farther will have a disparity that is smaller, that is what, in fact gives you a cue for depth because various pixels move in the image.

Now, that this again is not such an easy problem, looks like relevant if I do this and I am done. But then what really is the catch? The catch really, lies in the fact that you have a left image, right image and you have actually assumed that you would just translate the camera in

lane. Okay, so let us have maybe only along x. And so, in which case even knows where to search for.

So even know as to, as to what you need to, so you even know that this sort of correspondence it lies like analyze in the same row in the other image. But again, to be able to say that with a reasonable amount of the confidences is not easy because again you may have to run through some feature correspondences, you may have to use some existing sift or something in order to be able to tell which feature is going where.

So, you need to be able to establish this correspondence. And as you can imagine, it needs to be a very, very dense kind of correspondence, then only you will have a depth map that is going to look dense. So, all of that is not easy. So, I am saying that it is not like Stereos, it is all done and solved and all, so even there are open issues (( ))(2:36) they have been around for a long time.

And people are still figuring out ways by which you can actually improve this accuracy of stereo. And these techniques despite the fact that your laser based, rain finders and all that they have all come but still image based, depth map estimation still remains very attractive simply because of the spatial resolution that it offers and the fact that it is so very cheap and all that you need true images captured by a simple camera whereas there you have to spend those laser based equipments and are still very, very expensive.

I mean, you cannot really carry them around and so on. So that way, when they say people thought those things have come and then one day they will take over all the image based reconstruction techniques and all that that did not happen. And even today, people are still looking at using images to basically build a 3D depth map.

Now, despite the fact that this is an image processing course, that I am still talking about this because since you have understood how the image formation happens, both through a pinhole, so this stereo and all one thing that you have to realize is that whenever you use stereo, it assumes that all these images are perfectly focused.

That is an underlying assumption, which means that even if have a lens based camera and all, it does not worry about all that, just assumes that whatever you are capturing is absolutely in focus. Because the moment something begins to get blurred, then even to associate correspondence and all it is not easy, right? So, for example, in one image if this guy is sharp, and in the other image if that becomes blurred, then for us to be able to, associate this or get a

feature correspondence that will make it all the more tough. It is all these algorithms most of them right that for example, stereo or structure for motion or for photometric stereo for that matter, they all assume that what you have captured are all images that are absolutely in focus.

Now, what we want is both images ought to be focused. No, it is like this, when you say both are at the same depth, so what you are saying is say you are already going to say translating the camera and they are capturing two images. Now, the way you are actually constructing those whole equation is by assuming that you are doing an implant translation, and therefore capturing one image with respect to the first camera position, then you are capturing another that is the camera position.

Now, that is right based upon that, we had this equation and then which said that you know you can compute  $z$ , but the point is that what we are saying is but all of this assumes that say this basically assumes that you can get  $X_R$  minus  $X_L$ . But what is  $X_R$  minus  $X_L$ ,  $X_R$  minus  $X_L$  is supposed to tell you where this point has gone in the other image that is when you will get your  $X_R$  and  $X_L$ .

So, that is supposed to be disparity. Now, this disparity, how do you compute if, let us say one of the image becomes blurred? Okay, your center may still be alright, (( ))(5:35) still be blurring about the central ray. But your ability for example, if you I run any sift correspondence or any of the existing features, they all make a fundamental assumption that things are sharp. And, and especially right they do not expect when you walk from one image to the other, they do not expect the intensities to change because of blur.

In fact, if there is a change in intensity even because of illumination, something that will also create trouble but typically these are taken, within a very short time, right? So therefore, we it is not like you take one picture in the morning and then one picture in the afternoon in which case there could still be trouble. But then sift and all is supposed to be partially invariant illumination, partially invariant into post, not fully invariant. It is fully invariant to translation rotation and all, but not fully in variant illumination in post.

So, this partial invariance, what it actually means is that if for some reason if there is change in intensity, there could be trouble. But the trouble becomes more when one of the image let us say is actually blurred. For some reason maybe you just made it out of focus or you did not bother to focus image then now your ability to get this  $X_R$  minus  $X_L$  itself is in trouble.

Your equations are valid so long as you get  $X_R$  and  $X_L$  correctly. So, there is no issue with this equation per say. This all still holds, but the fact that your ability to get that  $X_R$  minus  $X_L$  is now in sort of a question, because you would not know where if this guy is smeared out.

So, to be able to tell that this is exactly that point will be hard. Okay, that is what I meant when I said that, you need both images to be in focus. Not only stereo, in fact structure for motion every one of them typically assumes that correct things. I mean, that is a reason right? Why what happened was when let us say somebody showed in those days, right if we were to show somebody a defocused image, then they would say first right, let me take an algorithm that will actually remove all the blur, so that I can operate all of this and little right did they realize that the blurred itself is screaming and telling that, here I am carrying a cue for depth, and you are sort of removing me first thing.

So, that is a way people tend to sometimes think. So, what I am saying is, so the cue for depth, in fact lies in the fact that that you are able to get  $X_R$  minus  $X_L$  accurately. And if you can get it for every point, then of course, you can have a nice depth map. But it does not always mean that it is an easy problem, especially if they take share.


Suppose, let us say you have a smooth texture somewhere here, I mean on a large region, then that large region will also exist here. Now, how do you tell which point has moved where because everything will look alike right. So, even if you think that things should not have moved too far, but still even within that region to be able to pointedly tell which point has gone where for example, I think last time I gave an example of something like this, I take an image of this and then let us say you incline it or something, so that there is a depth difference, you take another image then how will you match the correspondence, how will you get your correspondence, everything looks the same? Correct.

So, that is why when we say we expect image should have features we mean image should have some activity, there should be some changes in intensity, something should be going on, in which case we can get these matches, so we we know that there are things to match okay. So, the point is this right, so each one of these can serve as a cue and because now you know how a pinhole image is formed and you also know if you had a lens instead of a pinhole that what kind of defocusing can happen and things need not be blurred in the same way, there could be spacer in blur and all that you have understood all that now.

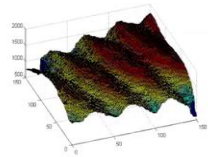
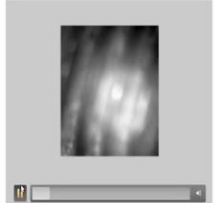
So, going along the same line what we wanted to do was understand that even for example, when you have a lens based right when you have let's say defocusing effects, or as it is called, defocusing or in general, see defocus means typically it is optical. When somebody says defocus, they mean typically an optical kind of blurring.

Otherwise if they want to be more specific, they will say actually motion blur. In fact, motion blur is also a cue for depth by the way, but at this point of time, I want to stick to something more simple which is actually optical defocus. Okay, so optical defocus, this is actually due to lens. So, here we are assuming that that we have a real aperture camera and we get an image that could be spaced very blurred. Now what I wanted to tell you was if I just pull this image out,

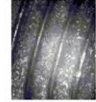
(Refer Slide Time 9:58)



### SFF using Tensor Voting



Reconstructed 3-D shape



Focused image

"Shape-from-focus by tensor voting"  
IEEE Transactions on Image Processing 2012

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

( Introduction to Shape from Focus )

Now, yeah, right. So, we can see that plane. Alright, so the other one was actually, I think I told you Sai Baba stays on a ring correct? I told you it is a ring. Now, these are just three wires by the way, these are just three wires that are being imaged under a microscope. Now, if I gave you anyone, there is a there is a whole bunch of frames that is being played and it is a video. Now, if I stop anywhere, we are supposed to stop here right.

Now, would you be able to tell, say now the whole idea is this right now we know that if there is a blurring, okay then actually means that the scene has to be a 3D scene, because if the scene was perfectly flat then we know that the lens would have either blurred it the same way all over or of course, it could have completely brought into focus right either of these two should have happened.

But the fact that we can see that some portions are probably a little maybe not in this frame level, let me go also here. So, it looks like some things are coming into focus and some things are still blurred and so on. So, now it means that sending out a cue that you are looking at some kind of a 3D object now, that means all points are not at the same depth from the lens.

Now, if I told you that from this itself given if I just gave you one image, and if I asked you right, is there a way that you can tell where each point is or what kind of a 3D object are you looking at? It is very tough because even if we had some kind of focus operator, let us say, I mean a measure if there was by God, if God gave us an operator like that, which we could run all over the image, and then it would tell, here is where you have a very sharp point and there you do not have such a sharp point, that if based upon some sharpness rate, if you could arrive at it, we would still not be able to do it because when you measure sharpness, the sharpness is not only a function of the degree of blurring that's going on, it is also a function about what is lying underneath.

So, for example, an image could have a very high activity in some places and could have very low activity elsewhere. Okay, so over a high activity and over a low activity, if you had the same blur, let us say, then what would happen is the high activity place, if you try to measure, you may still end up with an activity that is still considerably high, because of the fact that because operator that you are trying to use is only going to get as a measure, something that will tell how active is his region, and the fact that underlying that there is a lot of activity going on.

Whereas in some other place, the underlying activity itself is less. So, if you superimpose the same amount of blur, you would not know that the blur is the same, because you would think that this guy has less blur, whereas the fact is, they both have the same blur but then just that the underlying image has more activity here and then less activity there.

So, is your ability to be able to tell which is less and which is more blurred and all with a single image it is very, very hard. So, which is the why let us say the people go for actually multiple images. Now, if you see that, that is why you have a stack. You can even do it with two but then this is a technique that is called shape from focus okay and I want to at least introduce that to you as of today, so that you at least understand that.

Now, what you are asking is something like an inverse problem, till now what we had was we had I gave you a 3D scene, I said that right? This 3D scene will induce this kind of blur on


the image plane. And therefore, and if you assume that the image is of a certain  $(\cdot)$ (13:27) or something, then you could superimpose that. So, then what you get at the output is a blurred image. So, that is like a forward problem, I know everything, I know the 3D scene, I know what blur it is going to introduce, all that I have to do is get your image for you.

But now what we are asking is what is called an inverse problem. We are asking, given the image can you actually tell me something about the scene itself, about my 3D scene, and then that is where we found out that with stereo if you had two, you could still do. Even the blurring with two you could still do actually, because when you have two, there is some kind of relative blur that comes into notion now, because you have something, I mean if you have a reference with respect to which you can do, you can say something about less blur, more blur, it is okay, because that reference is unchanged, right?

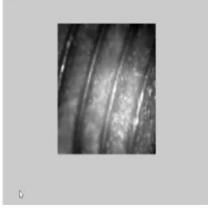
You have two images, and they both correspond to the same scene, and one is less blurred one is more blurred that relative degree you can still sense but this one senses more. Okay, this is not just one or two frames, I will come to that one, if we have time I will talk about the other one which uses just as because that was motivated by the fact that stereo users too so why cannot you do it with just two if you had blur also.

But let us go to the other one right, this one is more intuitive, this one is more reasonable to appreciate because so here if you see right so what you are seeing is a whole, this is called as stack, you have a stack, you have a stack of frames. Now, what things do you see in this stack? I mean the whole idea is that this stack is supposed to convey you information about his underlying object which anyway I plotted there, so there is some algorithm that is run on this stack, which gives a 3D shape of the object which is how it is supposed to look I mean there is a wire and then there is a dip and then there is again a wire, there is the dip and so on. Now, how did it sense it is what we want to know. Now prior to that I want to ask you what do you see in this image? You tell me what are all the things that you see?

(Refer Slide Time 15:25)

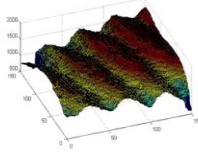


### SFF using Tensor Voting




Stack generated using a microscope

Click on the image for video



Reconstructed 3-D shape



Focused image

"Shape-from-focus by tensor voting"  
IEEE Transactions on Image Processing 2012

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

( Introduction to Shape from Focus )

I can repeatedly play if you want. One thing is that no single frame appears to be completely in focus is that correct? Nowhere Can you can you pick a frame which looks like all in focus. If something is going on somewhere, if something is getting blurred then something becomes focused. Something becomes focused then the other thing is other thing is getting blurred.

Other than that, what I wanted to tell you is, now if I tell you what is going on now you tell me, it is like actually having a lens at the top which is a microscope, right? It is looking at this object through a lens and there is a what you call a z stage, so z stage is one that is being moved.

So, you have kept this object on this so this is just a small, like three wires, and you kept them there. And then basically what you are doing is you are trying to you are just you are probably moving it upwards moving this state downwards and this guy remains right there. If I do something like that, something should happen. Now, what should be that that should happen? There should be some scaling effect that you should probably see right? Imagine no, right I am going like back and forth. I am going front I am coming closer or I am going farther off, I am translating along the optical axis.

When I do something like that, you would typically expect that something should get magnified. I mean what you seem to think this magnification is actually a parallax, but to our eye, it is not so clear so we seem to think that there is some scaling going on, but in simple terms, it looks like some scaling should have happened. But you see here right, this is it looks like every point is sitting right there, nothing is moving.



Now such as this one is called really a tele-centric lens. Now, we will not go to the details of the study centric lens, but I will kind of later on revisit this, but right now. So, our idea is to motivate shape from focus, we call it as an SFF, it stands for Shape From Focus.

(Refer Slide Time 17:18)

NPTEL

Stereo : Disparity map

$$z = \frac{f \cdot b}{x_r - x_l} \text{ Disparity}$$

SFF Shape from focus  
Defocusing (optical) / Motion blur

↳ Due to lens (Telecentric lens)

Shree Nagor (Columbia)

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

(Introduction to Shape from Focus)

So, what this means is that just as we had shape from mix, so shape from a disparity was earlier, now we are looking at shape from focus as a cue. So, this uses what is called a tele-centric setup. So, in fact, most of you are optical microscopes and all about come with this kind of a tele-centric setup. It is called a tele-centric lens what is tele-centric lens simply means is actually this is used in all industrial applications.

So, they come because what they want is if you keep something, if you keep an object like this, if you have a lens here and then you have an object here, some object here and then you see an image on the image plane. Now, though, if let us say later, if they want to match this object with something and they keep the object slightly away, they do not want this image size to change.

They do not they do not like some magnification going on and so on just because right you have moved the object a little farther or a little closer. So, in order to account for this what is normally done this in kind of a tele-centric setup what is done is you introduce an extra aperture okay at exactly focal length away from the lens center either on this side or on this side depending upon where introduced you get image side tele-centricity or what is called object tele-centricity. That is okay, we do not have to go to the details, object side means that you can move on the object side and then there is no magnification.

Image side tele-centric means that this image plane can be moved and the object can be right there, the image can be moved and then you still will not see any kind of magnification. There simply and there is more from an industrial perspective where people like a setup like that because then they do not have to worry about extra factors and factors like this, which they feel that are unnecessary to kind of deal with. But then this is exactly the point that share from focus exploits.

And by the way, I hope I told you that this method right is actually attributed to one Indian, his name is Shree Nayar, okay so that way you should be actually proud of the fact that our own person did a lot of work in this. So, his name is Shri Nayar at Columbia. So, his father, by the way was his grandfather was a freedom fighter I believe. So, it was he who came up with this idea, and he said that if you hired really a tele-centric setup.