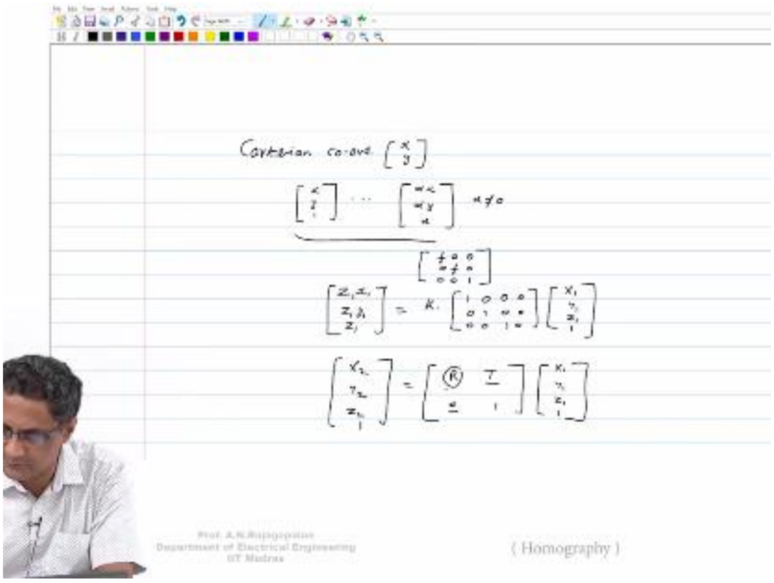
Image Signal Processing

Professor A N Rajagopalan
Department of Electrical Engineering
Indian Institute of Technology Madras

Lecture 11

Homography

(Refer Slide Time: 00:16)



You are right, I had said that you know we are looking at this homogenous coordinates and so if you have a Cartesian thing, if you have a Cartesian coordinate xy then its homogenous coordinate is the simplest thing would be x, y, 1, but then you can have all of these I mean any alpha x, alpha y, alpha, alpha not equal to 0, they are all, are all (basically) they are all equivalently represent the same Cartesian this entire thing which is why we call them as homogenous coordinates.

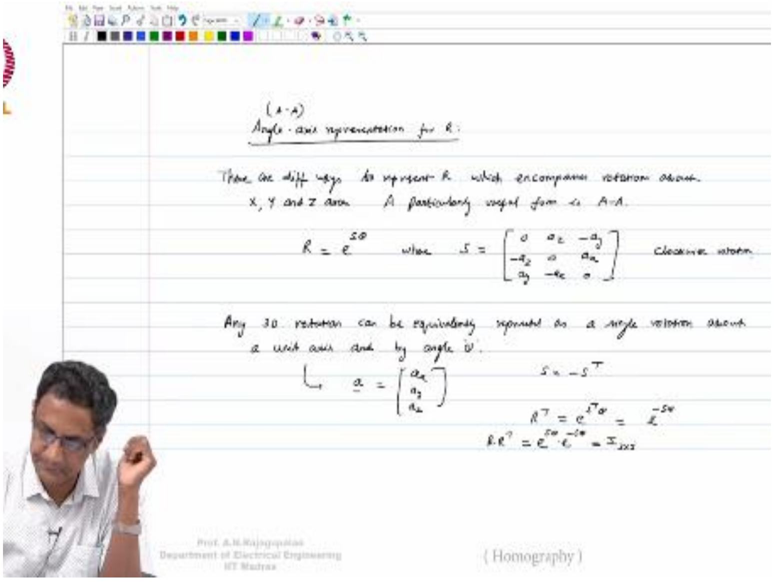
Because in the Cartesian space they all map back to the same point x, y. So, in SVS problem when we are trying to do a perspective this one a projection and we wanted to have a relationship which is linear between the scene point and your image coordinate. So there we took alpha, since alpha can be any, this is valid for any alpha not equal to 0 so we actually took alpha to be z1 which is the capital Z coordinate for the point P with respect to camera center o1.

So we did the following we said z1 what was it x1 then we said z1, y1, z1 is equal to let us call this as K where K is this matrix f 0 0 it is a 3 cross 3, 0 0, 0, 0, 1 where f is the focal length of the camera and then at 1 0 0 0, 0 1 0 0, 0 0 1 0 multiplying and then again the 3D coordinates expressed in homogenous form would be x1, y1, z1, 1. So in a sense using this homogenous coordinates instead you are able to now to create (()) (2:19) why linear?

Of course anytime you have a linear transformation that can explain something you would always go for that rather than something that is nonlinear. Linear has its own you know that linear has so many advantages over anything that is nonlinear. So, now you can also of course pull out z1 if you wish and then the way we kind of go back z the first two coordinates divided by the third one and that will take you to the Cartesian coordinate.

And then I also said that you actually so now if you were to go from o1 to o2 and suppose with respect to the o2 this point P that is sitting on that plane has a coordinate x2, y2, z2 then clearly we can write x2 y2 z2 1 in homogenous form this will be your R, I think I used the capital T for the camera motion then 0 then 1 being a scalar then this would have been x1 y1 z1 1. I think this is where I thought we had stopped. Now going further, now this R we would like to have an interpretation for the R which is of course a 3D rotation matrix which is actually a 3 cross 3 matrix and there are various ways to interpret this R.

(Refer Slide Time: 03:43)



The one interpretation that I feel is very handy and which is easy to understand is what is called angle axis representation and you will understand why it is called so, representation for R. This angle axis kind of, so let me just write this down there are different ways to represent R which is your 3D rotation which encompasses rotations about x, y and z axis. A particularly useful representation this is angle representation, a particularly useful form is AA angle axis, suppose I call this AA angle axis form.

So wherein we write R to be equal to e raise to s theta where this is a matrix exponential now it is not your regular exponential e power s theta where s I will explain what these seems to be. S is 0 az I will tell you what each of these entries are ay minus az, 0 ax ay minus ax 0. So what this actually means is that when you have I mean and there is a fundamental sort of a theorem which says which is actually due to Euler which says that any of these rotations that you kind of do in 3D they can all be equivalently represented, any 3D rotation can be equivalently represented as a single rotation about unit axis that you have to properly choose about a unit axis and by an angle theta.

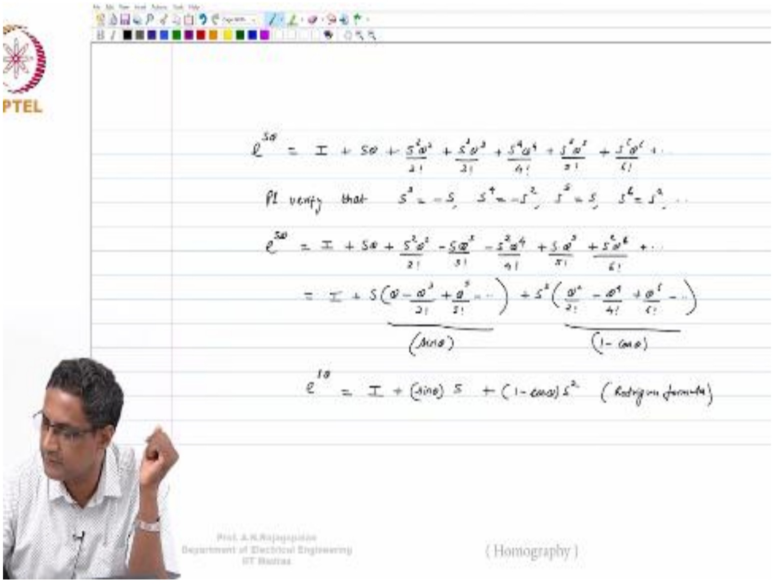
So this angle theta, this is theta that is sitting here so basically theta is a scalar, but s is a matrix and this unit axis is but a whose components are ax ay az and there are rules that allow you to go between rx, ry, rz composition and this r is equal to e power s theta and all I do not want to go into those details, but just it suffices to know that any sort of 3D rotation that you do in the angle axis form it can be equivalently because when you take this form then this is actually easy to sort of interpret all those in plane rotations.

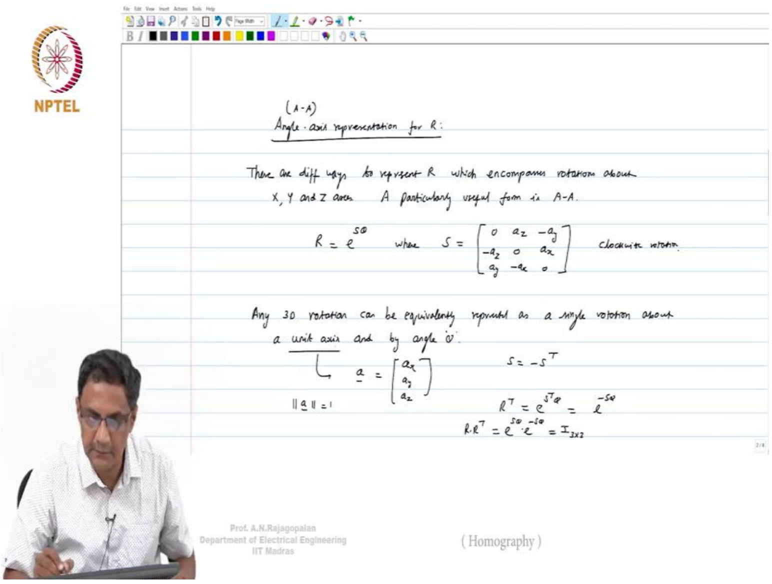
And anything that you want from shear to whatever you want that you can actually equivalently think about what should be the camera motion in the sensor, what should be this video rotation matrix and so about what axis we should be going say rotating clockwise and by the way this is for a clockwise rotation.

And as you can see right clearly s is actually (()) (07:52) metric matrix, s equal to minus s transpose because it needs to be that way even otherwise of course one is you just do check that by observation that s is equal to minus s transpose other way to argue is that because R should be orthogonal.

From here R transpose should be e s transpose theta, but s transpose is e power minus s, not this, e power minus s theta and therefore RR transpose should be e s theta e minus s theta which is equal to the identity matrix. So I told you this is a matrix explanation it is not 1 give you the identity matrix which is of size 3 cross 3. So this s is such that it is (()) (08:38) metric, it has to be.

(Refer Slide Time: 08:48)





Then going further on what we actually do is when we look at, when you look at e power s theta, so if you do a Taylor series expansion for this guy what you will get is I plus S theta plus S square theta square by 2 factorial. I am going to write down a few terms this is deliberately I am going to do this you might wonder why I am writing so many terms, but I want to plus s power 4 theta power 4 by 4 factorial this is power 5 theta power 5 by 5 factorial okay I will stop here s6 theta 6 by 6 factorial and so on.

Now this s and also note that this guy has note down this a has actually a unit norm so norm of a is 1 because I said unit axis. So this norm of this a is 1 if you use the condition that those norm of u is 1 then I leave it to you to show as a small exercise please show that I can do this here, but then let us not waste time. Please verify that s cube.

What do you think s cube might be equal to, anybody wants to make a wild guess? Is equal to minus s you can check this out, but use the fact that norm of a is 1 because you will get terms like ax square plus ay square plus az square and all you will get and you should equate that to 1 then if this is so then what will be s power 4 then minus s square because s power 4 is s power 3 into s, s power 3 is minus 3 so minus square.

S power 5 will be s power 4 into s so minus s cube, but s cube is minus s therefore this become plus s, s6 will be s square and so on, we can go on, go on and on and on. So now if you sure to, if you have to say plug that into this Taylor's expansion for e power s theta what do you get, you get I plus S theta plus S square theta square by 2 factorial.

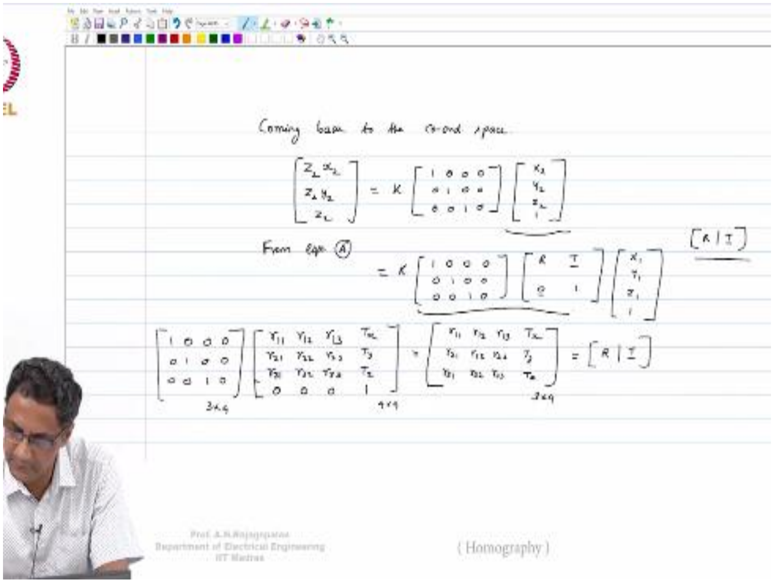
Now in place of s cube we will substitute minus s theta cube by 3 factorial, in place of s power 4 we will substitute minus s square theta power 4 by 4 factorial s power 5 is plus s theta power 5 by 5 factorial plus s square theta power 6 by 6 factorial and so on. It may be a minus after that. Now if you club all the terms that contain s and then club all the terms that contain s square.

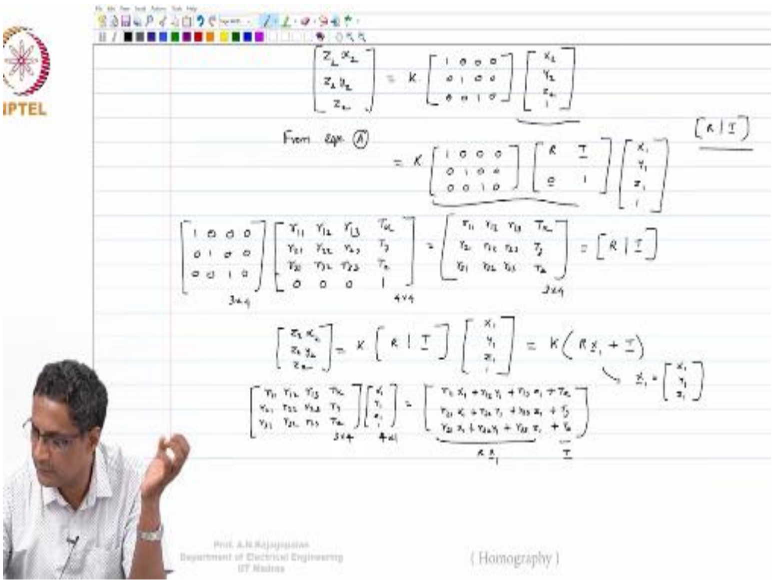
Then anyway, theta is a scalar so we can multiply either way does not matter for the right or the left really does not matter. So I plus S let me multiply from the left so theta minus theta cube by 3 factorial plus theta power 5 by factorial then it will be minus theta power 7 and so on so it will go on plus s square if you take you get theta square by 2 factorial then minus theta power by 4 factorial plus theta power 6 by 6 factorial minus and so on.

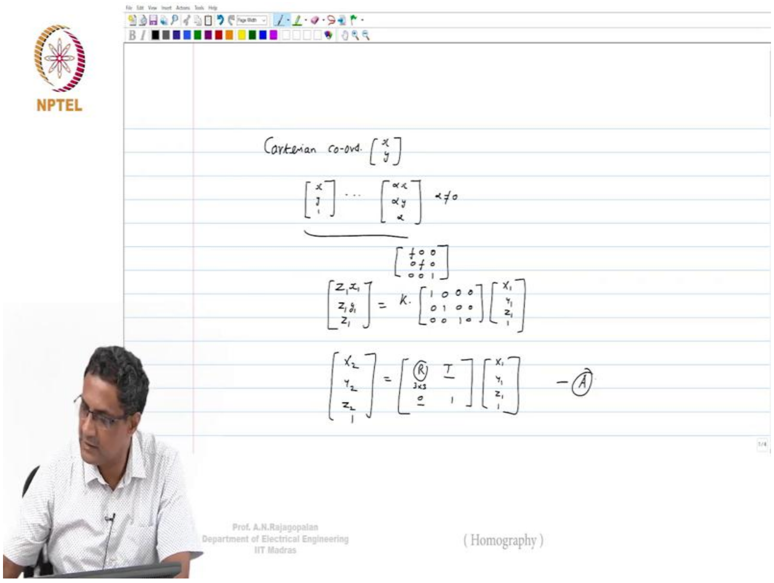
Now if you see these two guys you can relate them to something this expression is sin theta exactly and what is this? This is also something you know, it is not cos, it is 1 minus cos, it is 1 minus cos theta, so this is 1 minus cos theta. So this e power s theta has a close form expression which is I plus you can multiply whichever way you want I mean sin theta into s. So that means that every element that matrix will get multiplied by sin theta plus 1 minus cos of theta into s square.

Again s square is something that we can always find out given the entries of s. So 1 minus cos theta is again a scalar that will multiply every entry of a square in that matrix those are called actually Rodrigues formula. So, and as you can see if your angle happens to be 0 theta is 0 then of course automatically e power s theta becomes y and therein and all you can actually verify quickly. Now I will going to say, come back to this later when we actually want to like I said you have a camera motion, you have a scene and then what kind of an image will you get when you do that kind of camera motion that is something that we will see using this R and other things.

(Refer Slide Time: 14:10)







Now coming back to your homogenous coordinates, coming back to the coordinate space that was only an aside you just know as to how you can interpret R. One of the simplest ways to interpret R is like equivalently expressing it as a rotation by a clockwise or counter clockwise accordingly that s should change of course for clockwise rotation about a unit axis suitably chosen.

Coming back to the coordinate space now if you kind of go back to that figure we had a point P and then we have move from o1 to o2 and let us say we said that the coordinates of point P with respect to o2 now becomes x2 y2 z2 which then means that the image coordinates will then by just as we wrote for, when it was when the camera was at o1 similar to that we will write the image coordinates is z2 y2 z2 and is equal to K which is the, which is at focal length matrix.

Multiplying 1 0 0 0, 0 1 0 0, 0 0 1 0 and this multiplying x2 y2 z2 1 and then the whole idea is that we want, see eventually our idea is not to that is not to relate scene to the image our idea is to relate the two image coordinates. What we would like, ideally like to have is on this right hand side something that involves small x1, y1 and on the left we need small x2, y2 so that we can say that the two image coordinates can be actually related by a linear transformation.

So now what we can do is we can go back to that expression for x2 y2 z2 1 which I wrote down from that here. So the nice thing about this thing is that the established business you do not have to write anything again. So if you see this equation let me call this A. So from this equation what we can do is so from equation A we can actually substitute for this guy and then get K 1 0 0 0, 0 1 0 0, 0 0 1 0 multiplied by this was what R, T 0 1 and then x1 y1 z1 1.

We simply substitute for the camera motion. Now we know that T is 3 cross 1 vector representing camera motion. So if you actually multiply this you can actually quickly verify this, this will come out as R, T next to it if you want you can actually quickly do this. So let us say r is r11 r12 r13 r21 r22 r23 r31 r32 r33 and then you have your Tx Ty Tz those are these R and T.

And this guy is like 0 0 0 and this is the scalar multiplied this with x1, y1, z, no multiply this with 1 0 on the left, multiply this guy with 1 0 0 0, 0 1 0 0, 0 1 0 if you multiply this you will get so first row, first column so you just get r11 r12 r13, this is what, this is 3 cross 4, this guy is 4 cross 4 so you should land up with a 3 cross 4 r11 and then Tx and second one will be r21 r22 r23 Ty, third one will be r31 r32 r33 Tz, this is a 3 cross 4.

So what we normally do is we do not write like this we simply say this is a matrix R, T that is how we write. So it simply means that there is a compartment you know R is sitting and then next to it T is sitting as a column 3 cross 1 that is what I wrote here. So this will reduce so these two when you combine them together this will give you RT then, now let us just go up so what this means is that so we have z2, x2, z2, y2 and z2 to be equal to K.

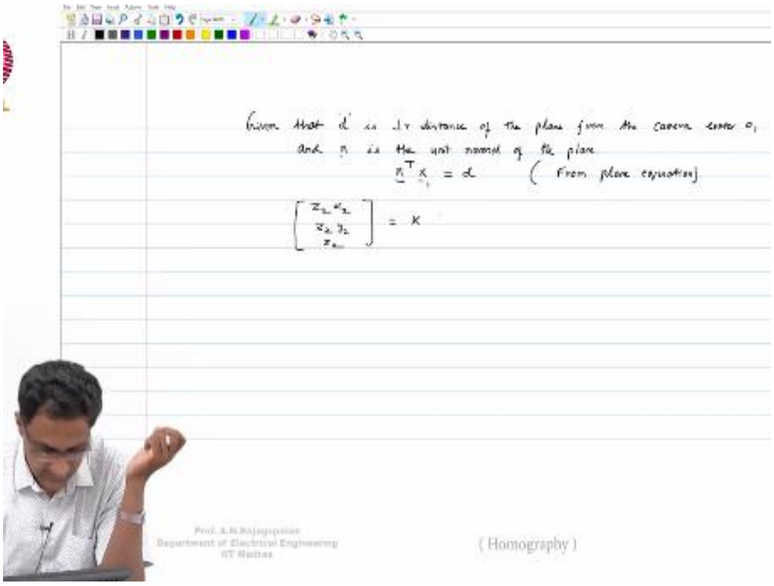
And then we have R, T because these two we have merged now and we have what is x1 y1 z1 1. Now if you actually multiply these two you know the form of this R slash T I told you how that looks like that kind of looks like this. If you actually multiply this with this can you get a guess what will happen, what will you get? If you want you can try it out or do you want or shall I just leave it for you to kind of check this out.

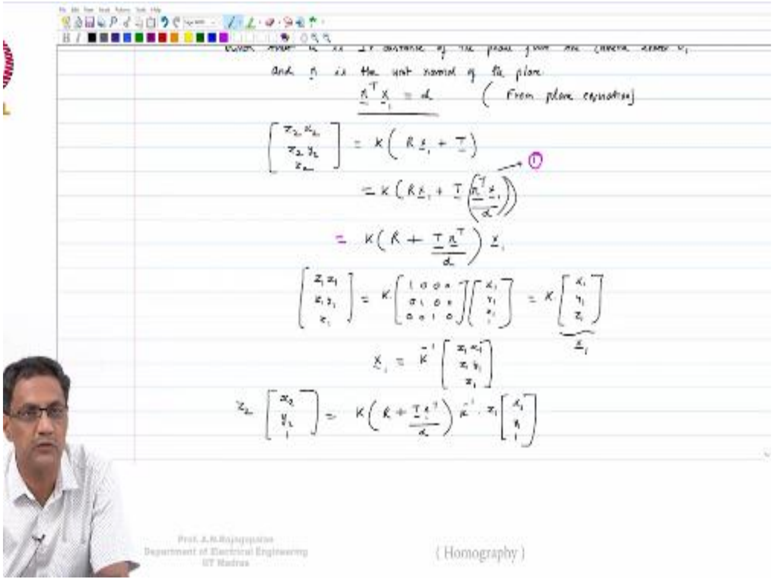
We will get something like K, R I am going to write this x1 plus T or else we can just do it of, it should not take that long r11 r12 r13, r21 r22 r23, r31 r32 r33, Tx Ty Tz multiplied with x1 y1 z1 1. So this is 3 cross 4 this is 4 cross 1 and so we should get the 3 cross 1 you want to quickly verify this so let us do that r11 x1 plus r12 y1 plus r13 z1 plus Tx, then r21 x1 plus r22 y1 plus r23 z1 plus Ty, r31 x1 plus r32 y1, r33 z1 plus Tz which is but rx1.

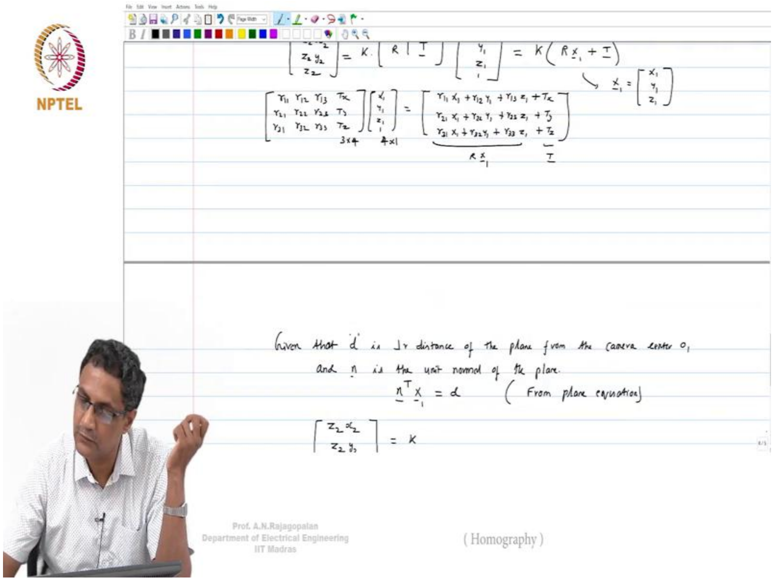
Now this x1 mind you is x1, y1, z1 does not have the 1 in it. So that is why I wrote it as x1 with an underscore just to repeat the fact that, reiterate the fact that this is not x1 y1 z1 1, it is simply the vector x1 y1 z1 because as you can see here if you multiply all that you have is x1 y1 z1 and then you add your translation Rx plus T so T is Tx, Ty, Tz. This matrix is but Rx in this manner if you write and this is your T so we are just adding the two and that is how you can write this.

Now until now you would wonder that really we have never used made use of the fact that we are kind of looking at a plane, but it looks like there was a point in the scene P which had a certain coordinate when you were in o1 and then you moved the camera and then you went to o2 you got another coordinate. Nowhere have you used the plane equation really because now is the time to use the plane equation because the whole idea is that we want to be able to relate image coordinates and not really scene coordinates. So now comes the role of your say plane equation.

(Refer Slide Time: 22:45)







Now given that so if you go back and check in that figure that I had drawn what we had said was given that d is a perpendicular distance, given that d was the perpendicular or d is the perpendicular distance you go back and check I would have actually mentioned this distance of the plane from the camera center when it was at o1 by the way when it goes to o2 it can change the d can become something else of the plane from the camera center o1.

And then we also had actually a unit normal and n is the unit normal of the plane. Given these two facts we know from a simple plane equation that if a point P lies on that plane and if it has whatever xyz has the coordinates then we know that it should satisfy the plane equation and transpose, I will write this is x1 is equal to d because the P has coordinates x1, y1, z1 the homogenous form is x1, y1, z1, 1, but the plane equation is what, n transpose x is equal to d.

This is from, I mean this you all are familiar with this is simply the plane equation from the plane equation, okay here is where we are actually making use of the fact that this scene consist of the plane until those point we never made use of anything. So we know that n transpose x1 is equal to the scalar d. Now you walk back to that equation that you had let us say previous page which was z2 x2, z2 y2, z2 and then you had K and what was that Rx plus T or something.

What did you have there? So you had z2 x2 k into Rx plus T, Rx in the sense that we wrote this is Rx1 plus T. So now what we can do is we can actually play a small trick because in order to be able to equate the image coordinates what do you think we should play as a trick, what should be the trick? Given that this is valid for any point P on the plane, so what should we do?

We should just take K Rx1 plus T n transpose x1 by d because that is a scalar n transpose x1 is anyway a scalar that is d. So all that you are doing is multiplying d by 1, but then that tricks helps you to pull x1 out now because otherwise you have an x1 sitting here and then T is sitting there but you want to bring in your image coordinates. So you just, so this is n transpose x1 which is d by d.

And hint of the dimensions there is not anything wrong because even if you try to look at T well the ideally rate, the way we interpret this is like n transpose x1 by d. So this quantity is 1 this is just 1 so then what we can do is we can actually pull of course now we had to pull x1 on to the right side because it is now a vector so R plus T n transpose by d into x1 because x1 is common to both the terms inside.

So now we have at least reached a stage where there is x2, y2, z2 on the left hand side and then we have x1 standing alone now. This is still a scene coordinate. Now if you go back to that equation that we wrote that z1 x1, z1, y1, z1 what did we write this as the focal length times and then we wrote this as 1 0 0 0, 0 1 0 0, 0 0 1 0 times x1 y1 z1 1 this was one of the first equations we wrote.

And then this also means that if you multiply it this will give you K times x1 y1 z1 this is but your x1, the way I have been indicating x1. And therefore and K is simply f 0 0 0, 0 0 0 1 therefore x1 from this equation is k inverse z1 x1, z1 y1, z1 and please note that right it does not mean that one can go to a scene coordinate, one can go find out where that point is given an image point.

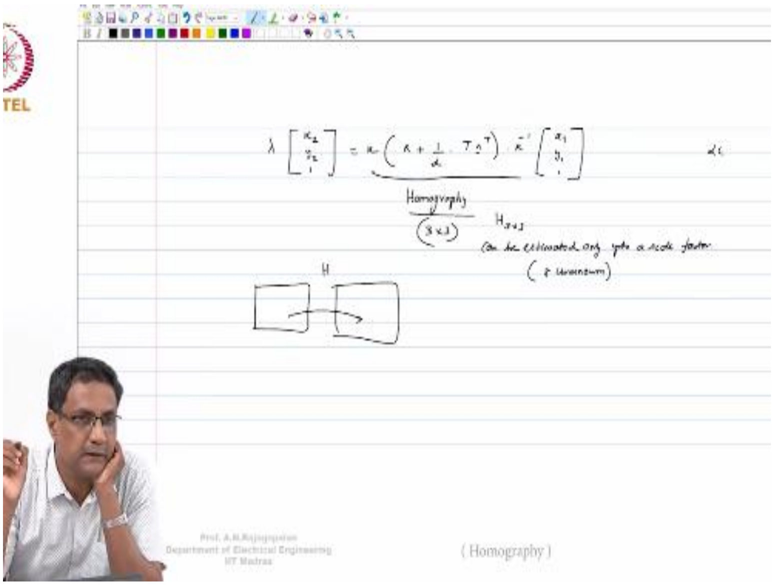
It is never possible because when that point it maps on to the image you cannot go back and tell where was that point, but here you are assuming z1 knowledge then you can write this without z1 knowledge you cannot write this. So it does not mean that this inverse means that given an image coordinate you can go to the scene point no we are assuming knowledge of z1 which in the reality we do not know what is z1.

But for us for kind of completion of equation this is fine. One should not interpret this equation as something that takes you from the image coordinate and straight to a 3D point that is what the whole problem about estimating 3D nature of a scene from images. Why it is such (()) (28:38) problem because my point here maps to as let us say an image point here, but even the one behind also maps there, this the whole ray, so where do you stop?

So you cannot directly pinpoint and say that here is where that point is because entire rays mapping to that point, but here that z1 we are assuming this equation the z1 is sitting inside the image coordinate that is why we are able to write this. Anyway I mean you do not have to you may not even think so much about it. I just wanted nobody to get confuse at one can go back to the 3D point from the image coordinate you cannot, unless you have more than one view or something.

Now here, so now we can go back and this is substituted for x1 here which does not means that now let me just pull out z2 and write this as x2 y2 1 and we will have K into R plus T n transpose or we typically write this as 1 by d times n transpose and then x1 I am going to write it as k inverse again I will multiply this by the scalar z1 and I have x1 y1 1. So now what you have is some x2, y2, 1 in the homogenous form and then it is getting, and then it is being related to the image coordinate of the first view which is x1, y1 written in a homogenous form.

(Refer Slide Time: 30:03)



So which then means that we can go further and write this as some lambda where lambda is some z2 by z2 and you have x2 y2 1 is equal to k times R plus 1 by d T n transpose K inverse x2 y, sorry, x1 y1 1. So now, note that Tn transpose is going to be a 3 cross 3 matrix because T is a column vector Tx T by Tz n is some nx and y and z this is unit normal you multiply those like taking an outer product.

So this will be a 3 cross 3 matrix scale by d and then added to R which is also a 3 cross 3, K is 3 cross 3, K inverse is 3 cross 3 so this entire thing this is what is called a homography and as you can see this we could not write in the form that we wrote earlier we had simply xy and then we could relate in plane rotation, in plane translations all with simply 2 on this side we had only 2 cross 1 vector.

But the moment you go higher up when you try to talk about a projective transformation you resort to homogenous coordinates and only the homogenous coordinate space can you actually, can you relate the two through a linear transformation and clearly homography it is simply a matrix which is a 3 cross 3 matrix and because of those lambda which is a scale factor sitting there this homography, let us call this as H, it is the 3 cross 3 matrix can be estimated only up to a scale factor.

Which also means that it does not have actually 9 unknowns because you can only estimate it up to a scale factor. It is like saying that you know if I had a vector ax plus let us say by plus cz and suppose you said that I can only estimate up to a scale factor then it means I can only find as alpha times the whole vector.

So I am obviously divide the last I mean your z coefficient whatever it is let us say it is alpha c so I am going to just divide the whole thing by alpha c so then I will have one coefficient for x one for y and then z will have just the unit as coefficient. Similarly, this homography think of it as you put up as a kind of a vector, whether you look at it as a matrix or vector it does not really matter, you can look up on it as a 9 cross 1 unknown vector or you think of it as a matrix.

Ideally you should think of it as a matrix and because it can be only estimated up to a scale factor so it has effectively 8 unknowns and we call them as unknowns because when somebody gives you these two images coming from two different views of the camera what is normally asked is what will be that homography that will actually relate these two images, that is because, that is the problem that we actually typically end up solving.

Because nobody tells you directly that homographic matrix, people will only typically give you two images where people will give you two images and they will say that these are images taken of a scene from a camera that has whatever moved in some manner we do not know how it moved. Now find out what is that homography which actually relates these two. So what is that H that takes you from here to here.