Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-71

We have already started what is called structure from motion and structure from motion is more about reconstruction, a 3D reconstruction, just to give a quick recap. The 3D reconstruction which means that you have a scene and you have these cameras which are probably looking at it from various viewpoints and the idea is that each of these x, y, z that you have the scene coordinate and you want to be able to reconstruct the scene, okay that is why we call it a 3D reconstruction and the idea is you can take a camera handheld or whatever you can go around a 3D object, capture multiple images with that and use all these multiple views in order to build the scene structure, that is why it is called structure from motion and motion refers to the camera motion and structure refers to the scene structure, right and this typically comes under the case where there is a certain things are not known in the sense that the earlier case and all that we saw was fairly simple because we had knowledge of the baseline and all that and therefore we could easily compute z and so on. But now we are more interested in kind of 3D reconstruction and using, I mean you can fix any one of these cameras is actually a reference view and then you have to sort of, so it involves, right motion involves the pose of each camera which means that the r and t, right corresponding to each camera you want and the scene structure refers to x, y, z, right that you want, okay. So basically that in a sense is what is involved in structure from motion and this is slightly, this is definitely more, this is definitely harder than the stuff that we have seen till now and in the last class I just showed, I just took up a two view situation because that is the simplest and today's class hopefully right we will do what is called multi view SFM and in the two view SFM you can of course initially compute the fundamental matrix, then you can use the calibration matrices in order to arrive at what is called the rate essential matrix which is a decomposition of translation and r, right Tx into r. And in the last class I showed that there is a way to actually decompose this, there could be multiple choices for the Tx and r but there is one choice which gives reconstruction such that the points come in front of the camera, what is called a chirality principle, so one uses the chirality principle to sort of say which of these configurations is acceptable, there is only one configuration for which the points project and the front of the camera or the points are in the front of the camera and that is how we showed that you can arrive at Tx and r and once you know Tx and r then you can go and do a triangulation because then a projection matrix is now available, you have the camera intrinsics, you have the r and T therefore you can actually triangulate and we showed last time that given one point correspondence right, you get 4 equations, linearly independent equations and those you can solve using SVD in order to arrive at xyz, right. Now going forward right, we will talk about a few more things now along the way, now what if let us say what is the camera matrix, the intrinsics is also not available to us right, then it kind of boils down to estimating your pose or the P matrix right, what do you mean finally when you want to do triangulation you need the P to go back right, that is what we showed and the case when let us say right you do not have the intrinsics available to you, then one has to get a projection matrix from the fundamental matrix itself right, because you cannot go to the right essential because that camera, the camera intrinsics is not with you right.

So the first thing okay that we will see today is you know deriving a projection matrix from F okay, so projection matrix from F right, I am still in kind of two view situation okay, the multi view we will see how to do what is called batch processing and so on, so projection matrix from F right, so what we are sort of saying is that the intrinsics not available, camera intrinsics okay. So in a sense right without loss of generality right, what we will say is we will take the PL to be just I0 because anyway the intrinsics is not available with us and we will take PR to be in general of the form some mode over right, you can take it as AB, as you know this is a 3 X 4 matrix and right and this is like and this PR is going to absorb everything right, the intrinsics as well as you see extrinsics right, so this A and B will be, will have to absorb both and A is 3 X 3, B is 3 X 1 right. And there are actually that one can show that there is a, see the more relaxed you make it right in the sense that the more you say that I do not have this, I do not have that in this case right, we are now saying that I do not have the intrinsics, the more you go down that path then what will happen is the good thing is that you will still want to ask whether I can solve the problem to at least to some extent which is a good thing but at the same time certain ambiguities will arise right because you are kind of leaving the thing more and more open now. So there is not just one possible choice for A and B okay, there are several choices but one possible choice which is the most commonly used is this A right to be the A as the X of ER right, so ER is the you know right A B pole times the fundamental matrix, the fundamental matrix we are assuming that you have an 8 point algorithm or something right through which you are computing F okay, so F is this 3 X 3 right that we have with us that we have computed using point correspondences and this is of course 3 X 3 right, this is ER expressed as a matrix and that is a X product form and F is a fundamental matrix and B itself is ER which is 3 X 1 okay, one possible choice right is this and why does this make sense right, why this makes sense is because of the fact that right now suppose let us say suppose we go ahead and so one pole, so let us say that let us say that let X ~ right as usual let us say it is X1 right where this X is usual X Y Z and okay and then let us kind of look at look at you see X ~ right which is my PL acting on the capital X ∼ and PL given that it is i 0 right, so this turns out to be X itself right which is X Y Z and this X ~ dash which is supposed to be PR acting on X ~ but my choice of PR itself is such that right it is like of the form ER X F and then I have the B which is ER and this I am going to multiply with X ~ which is like X Y Z 1 right and this we know can be written as ER F into X right + ER right.

Now if already if our choice is valid right if our choice of A and B is valid then what then we need to examine whether X ~ dash what happens to this right X ~ transpose of course, right we want to examine what is this number equal to, so then this will be equivalent to ER X FX right let us say FX what is this X ~ dash transpose right + ER the whole transpose and then FX ~ but X ~ right from here is X, so this becomes FX right or we can write this as ER FX well yeah one other thing that we can do is we can in fact write this as actually a X product of ER with this vector FX right and this whole transpose right is this guy and if you multiply FX right then you will get FX here + you will get ER transpose FX right if you if you just expand it right and now we know that we know that right a X product when you take ER with FX and then if you take a dot product with FX that will be 0 right and this one if you watch out right this is nothing but F transpose ER the whole transpose times X right this is the same as this and we know that F transpose ER is actually 0 right this this we have already shown and therefore, right it is clear that this entire thing is 0 right therefore, what this means is that that choice of A and B that we made satisfies the the kind of fundamental matrix matrix equation which is this X ~ dash transpose FX ~ should be equal to 0 and then therefore, right that choice is actually a valid choice in fact, right this is not the only choice in general right I mean in general one can show that P R right in general can take this form ok what is like ER X F + ER some V transpose and then this can be some λ ER your B where λ is some constant and V is any any arbitrary 3 X 1 vector ok. So, in general right any of these choices is valid as far as P R is concerned. So, the point is right. So, when you when you sort of. So, when you go from less sorry more knowledge to less knowledge of what you know about the camera right it makes things more and more it takes makes things harder, but yeah, but then it is not unwieldy right you can still still still kind of do things and all of this right leads to what is called what are called structural ambiguities in in the in the say estimation of the you see 3D scene.

That means, when you reconstruct a 3D scene right there can be inherent ambiguities which arise depending upon what you know right. So, you can actually classify them as two cases one is sort of a calibrated camera case where in we mean what we mean is the intrinsic are known and then you can have another case where you say that it is uncalibrated right. So, both cases are of interest. So, we will see what what we mean by what we mean by structural ambiguities in in you say S F M structural ambiguities in in this S F M. So, what this means is that right when you kind of reconstruct a scene right it does not mean that it appears the same way that it appears to your to your human eye right.

Just like just as you know with respect to homography right when you applied a you know sort of a projective transform right on a on let us say x y 1 on an image on image coordinates we saw that you know your parallel lines need not remain parallel depending upon right what kind of a transform it is right. If it is most general which is actually a projective homography right then we saw that you know lines remain lines, but then parallel lines right need not remain parallel and so on. So, you get something like I do not know a distortion right you call it a distortion right if you want. So, similar to that right now when you are when you are doing a scene reconstruction which is like x y z 1 right there it was like x y 1 all image coordinates you know looking at like x y z 1 where x y z is the scene coordinate. And now you are expecting that that when you do a scene reconstruction then that scene reconstruction can can have structural ambiguity.

We call it structural because it is all about that is what we mean by structure right you are estimating the 3D scene itself. And therefore, something similar to that also happens here. So, it is nothing unusual right we have already seen it except that we saw it earlier in the form of a homography. So, it is no longer a homography, but the actions are roughly similar the implications and you know the kind of things that we saw there right can also happen here ok. So, that is the idea.

So, let us take the case 1 what is called what are called the what is called the calibrated cameras again right I am just I am just looking at looking at the yeah. So, let us. So, let us say that let us say that we have x ~ dash which is let us say p r. So, calibrated p r times x ~ right which also means that it up to a scale I mean whenever I write this write a projective we know that is all valid up to a scale right. So, I can always you know put in a λ λ not equal to 0 and say that it is still equal to this right it does not affect anything.

I mean this we are bringing in just to bring in some notions about scaling rotation and translation and all that. Now, if you know p r right. So, let us say p r given that this is a calibrated situation. So, we know that we can write this decompose p r as k r into r t into say into you know x ~. So, we can take actually k r to the other side and that gives you like k r inverse x ~ dash is equal to r t x ~ right.

Now, let us call just for ease right we will call this as some u ~ dash that is like r t x ~. And again it because only say homogeneous coordinate you know coordinates are involved that I can I will multiply this with some alpha alpha not equal to 0 and still go ahead and kind of write it like this. This I am kind of writing for a for a for a reason. So, now if you see right what you see is that what you see is that the coordinates right that I mean. So, this is all kind of going back to saying that based upon image data right what can I infer after all what do you have with you just images right.

You only only images taken of a scene that is all you have with you and in in in addition in this case you also are assuming that the camera intrinsic are known to you. So, typically in s f m it will be the same camera ok. So, you would not have like k l and k r and all I am just sticking to that because I am using a two view, but typically it is a same camera that goes around. So, there is nothing the k r k l and all it is one k that goes around, but because it is a two view case I thought that just call it k l and k r ok, but really this is nothing like k r and all. So, λ u ~ dash so, now the idea is that right.

So, if you if you think about it right there is a scale factor here and there is and there is a rotation right and there is actually actually a translation involved there right. So, of course, this comes from the right the camera pose and of course, there could be an alpha factor sitting there. Now, you can actually effectively write this as r t right nothing would change if I were to if I were to multiply it with let us say some H s which I call H s H s inverse H s which is actually a 4 X 4 matrix, but this is let us say a similarity transform. What does that mean? So, so similarity transform will typically mean that let us say you have something like suppose I call this some r dash and then some say t dash then I have a 0 transpose and 1 some 4 X 4 right invertible matrix which is a similarity transform right. So, why a similarity because your alpha r t you know that when you have scale rotation and translation and that is actually a similarity right that is what we saw in 2D also.

So, what this means is that you can actually have I mean right the x. So, so what this means is that right this remains a similarity transform. I mean if you actually compose H s inverse with alpha r t right that will that will still have scale right you know this one rotation and translation. Whereas on the right hand side what has happened is H s has gone and acted on actually right x ~ and H s itself is actually a similarity transform. So, what this means is that right H s can modulate x ~ in the way that it could scale it, it could rotate it, it could it could do a translation whatever right it could do all 3 and and right you will probably not see the actual x ~.

The actual x ~ is let us say corresponding to alpha r t in some absolute sense. But because of this ambiguity right there is a there is an there is an inherent ambiguity that I can throw in. I can say that you can have an H s inverse H s and still all of this is valid because image coordinates have not changed at all. See finally, I am going back to my image coordinates right that is that is where the correspondences are that is the only thing which tells me what I can do. And the image coordinates have not changed right I am still kind of looking at the same correspondence.

But now the structure is not the original structure, the structure is actually kind of modified now which is which is what we mean by a structural ambiguity. So, what it means is that in a calibrated camera right you can only estimate structure up to a similarity right. That means, if you get a if you see a structure right that looks smaller than what your what your actual actual feeling about the object is right it is possible right. It can be rotated it is possible right those that ambiguity can come in because you cannot stop it right. So, this is like an like an inherent ambiguity ok.

Because of the because of the fact that all this is coming because of the fact that you do not know the actual r and t you do not know exact pose right. See that is why it is structured when you say s f m the camera poses are unknown. See if I write somebody told me what was the r and t and all then it is all done right then there is no ambiguity. But because of the fact that the camera poses are unknown right this taking a camera I take a picture from here I take a picture from there I go there take a picture go there I take a picture I am not recording how much I moved and all I am not I do not know where I am, but I want to know where I am right. But to what extent you can tell where you are there is an ambiguity there right and that ambiguity for a calibrated camera right this is still a calibrated camera even for a calibrated camera.

So, right let me just write down. So, for this calibrated camera case the structure the scene structure right can be can be see determined up to up to a can be can be determined only up to a similarity transform. In the next class I will show you some some examples ok in this too just visualize what that means. So, so, so then it automatically means that if you have if you have an uncalibrated case then then surely everything should be even worse right which is what happens.

So, for an. So, if there is a mixture like. No, no that is all if you knew E essential matrix. No, no no no no no no no no no no no no no Which one? We did not know E right that. Ok no no no no no say that again no no yesterday was actually calibrated case ok go ahead.

Calibrated. For the caliber ok go on. We know that we are trying to estimate as a product of s and r. So, from that we get an estimate of the rotation. You get an get an estimate in scale the the I mean translation is basically right only up to a scale.

Ok. Right see that r see the point is right yesterday what we showed was among those four choices right that you made right depending upon any one of them right we saw that that is what brings the points in the front. But what basically right this is showing is that even for that choice whatever you make on top of that there is there is an ambiguity inherent ok. That means that means this this H S right will not will not make the points go back or something ok that would not happen. But then there is an additional ambiguity on top of the r and t right which we have found because see what this means is ok literally right what this means is position camera position and camera translation cannot be found absolutely. That is what that is what this effectively means position and translation cannot be cannot be found right absolutely there is always an ambiguity.

This is because of the fact that we are assuming assuming a totally sort of what you call a totally you know free what you call you know hands free situation in the sense that you do not have any other information. If somebody told you a little bit more for example if they said that there is one point in the scene whose depth I know or for example right if they said something like let us say from the camera right if there is a sensor that tells you how much you moved then then then all these issues will not come. If you just go by image data right if all that you have is image and then feature correspondences of the image and based upon that when you want to go back and and this estimate a scene structure then even the camera of motion can only be found up to a certain ambiguity because that is also as you saw right R T and then you have an H S inverse coming and multiplying that. So so really right it is not like R and T are absolute and similarly the the the scene structure is getting multiplied by another similarity transforms if you say right that is also not really absolute. So so really there is nothing like you cannot measure anything in millimeters now.

So you cannot say that the the object is 3 millimeters away or something. It is not absolute right so so all other so there is there is something more there is something trickier than this but I think if I say all that right then it will it will make matters you know I mean we do not have the time for that. I will I will probably pose it as a question later towards the end of the course or if somebody asks in the middle right at that time I will answer if somebody asks it then we will see. So case 2 right uncalibrated because idea is that whatever whatever H S inverse is doing right H S undoes it right and therefore, this ambiguity is there. So uncalibrated camera right so here so in this case right you do not know k L you do not know k R right so very intrinsic is not known.

Therefore what you have is something like this if you rewrite the rewrite earlier situation it will become something like P R into you see X ~ right because because we do not we cannot decompose P R now we cannot write it as k R R T and all because we just do not know and this P R where will it where will it how do we how do we get access to P R who gives P R? P R comes from what I showed right you get the fundamental matrix and from the fundamental you can get your P R except that there is no notion of intrinsic and all now right. Yeah intrinsic are not known so it is it is kind of it is a what you call you know it has got right everything inside it. So what is then means is that you can have something like this you know let us say an H inverse H X ~ where this is a 4 X 4 of course, invertible matrix a a projective ambiguity a a projective ambiguity. So what this now means is that right you can so this is still valid you will still get the same coordinates but now this is this of course, a 4 X 4 matrix now see you write what it is doing it is acting on say X ~ right and this is like what we saw with respect to a 2D case right you have like X Y Z 1 now X

~ is like X Y Z 1 and then you have actually you know listen a projective you know right ambiguity which basically means that means that means that you know H can take the form a I mean a dash b dash 0 transpose 1. Yeah because I mean right I mean if you if you if you kind of if you kind of see the see the structure of structure of P R ok P R is 3 X 4 right yeah in general I think I write H can be H can be a full blown H can be a H can be a full blown sort of you know a projective ambiguity.

So what I am saying is that you cannot interpret things anymore as translations rotations and all see I mean this I mean this a dash b dash it is the one right you know which is going to change you change your a and b in your in your P R right. So what I am saying is now there is no interpretation that this is a rotation and translation and so on this is just a projective ambiguity and what it can do is exactly what it did what I what I know this one a projective homography did for the image coordinates. So whatever a projective homography did for image coordinates something similar is going on here but with respect to a 3D 3D points now right. So this X ~ is like right X Y Z 1 right. So this X Y Z 1 so this so this so so what this means is that no parallel lines need not remain parallel so for example right if you if you had I mean right if you had two parallel lines in in the in the original scene right whatever you think that that is how you should have been but then after reconstruction you find that right they are they do not they do not seem to be parallel anymore that is all acceptable that is also an acceptable solution only there is nothing wrong with it because of the ambiguous nature of the model right but that does not mean that I mean you know it is all useless or something right just that given that if you have only access so much information you can only do so much the more you give me the more I can actually improve that is why there is something called stratified sort of you know reconstruction.

So what is stratified means is that a perspective sorry I mean a projective a projective and there is also this was an in between ambiguity right I am not I am not kind of right or no no talking about well actually actually right no no actually right this is this is this is more like an you see affine ambiguity this is an affine ambiguity H H can be full blown okay four X four this is affine what I wrote was affine okay so so there is actually a projective to affine to actually you know a similarity right in that order you can go projective is the worst right then followed by actually and I say affine yeah right he was right he was asking right should the last one be one need not be a full blown projective need not have that and affine has to have that and affine will have like zero transpose one okay because I will talk about what an affine model is but an affine model is such that a projection matrix it looks like it will be three X four but in the first two rows will have like p one one two let's say p one four and then p two one two p two p two four the last column will be like zero zero zero one so the p itself will have that form p itself will have that form and therefore right that forces the the H also the so in that case you can write it as H a inverse H a where you want if you want to say that it is an affine ambiguity you can explicitly say that so for example, I can I mean for an affine ambiguity right okay this is actually affine okay so for an affine this will become like p r H a inverse H a x right x ~ which means that in an affine ambiguity parallel remains will remain parallel a projective is the worst okay and and in an affine case this p r itself will have that structure so the p r right will look like this actually p one one p one two p one three p one four p two one p two two p two three p three four and then zero zero zero one that is the form it takes okay I mean if it is we are not talking about affine models and all there are actually various camera models okay and this ambiguity and all comes depends upon that's why when you say stratified reconstruction right what it actually means is that you can go all the okay let me just write that down so what you mean is when you say stratified right reconstruction reconstruction so what it means is give me additional information and I can go I will go from a projective to affine to similarity right and and how can I go right it depends upon how much information you give me right so if you tell that the camera model is affine then I can then I know that it's not say really I know a projective kind of an ambiguity that means parallel I mean will still remain parallel that much I can tell but but then right if you also if you also give me some other information so the information can come either from the scene in the sense that some even if you know one point in the scene for whose absolute depth you know then you can do a lot of things okay but this is assuming that you don't know right anything at all about the scene or you can say something about the camera model like I said it in this case if you tell me that it's an affine camera or something then then there is there is right certain things which you can do or if you tell me motion like for example they have some sensor on your camera that can tell me something about the camera motion any of this information can help me come from a completely full-blown sort of you know this one a projective ambiguity to something simpler okay I can then reconstruct my scene accordingly okay is this clear okay so so yeah so so so please note that is affine okay this is an affine ambiguity okay now now kind of let's kind of see multi view multi view right this is an.