Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-55

Okay, so we have already seen features, feature correspondences and all right, so we know now how to, so if you have an image, if you have another image then we know, we know when what are the, what are the key points that we can pick from image 1 and then we can also look for, look for points of interest in image 2. It could be any of those, any of those that we have studied right, Harris, corner or sift or surf or whatever and then we also know how to match them right, we have a descriptor so we know that this probably goes here and that goes there whatever it is right, so we know to match, now with this right, equipped with this knowledge, next question is what can we do right and we did actually motivate right, why we might need to do these correspondences and all, why we need to establish correspondences, let us go one step forward and ask us to what we can do right. So the applications are many and this geometry right, so the way we will actually, the outline right of this, of our approach to geometry will be like this, we will initially look at what are called, what are called 2D, 2D geometric transformations or 2D image transformation or just call them image transformations, transformations wherein we will assume that what we are looking at is actually a planar scene but we might wonder it is planar scene so interesting that means the first part will not have anything to do with depth or anything right, so it has nothing to do with depth, we will just assume that the scene that we are looking at is planar, when it does often, it does happen that often times we do look at planar scenes, the other way to look at it is something that is far away right, we can always approximate it as planar because if the change in depth within the scene is much smaller you know as compared to where it is from the camera, you can sort of you know merge all of them and sort of say that what I am looking at, for example if you are flying from a plane right, what you look down, when you look down right, you sort of approximate the, you may have buildings and all right which means that there is a depth difference that is coming up but then it is so small, it is negligible compared to the height from where you are watching right. So you can think of various situations even in close for example that wall is actually a planar right, planar wall it is not like I am very far away or anything, so you could have structures that are inherently planar in which case you do not need any assumptions about how far away you are but if you are, but if it is a 3D scene then yeah right, then maybe right you might just want to make the assumption that you are sufficiently far away so that you can model that as a plane. Now once you have this planar kind of assumption right, then it is like a flat scene right, when you say it is planar it is like a flat scene and on a flat scene right you might wonder what can I do, I mean if I have a flat scene but the interesting thing is you could do lot of things, in fact the panorama that you use in your cell phones right to build a wide field of view that comes out of the assumption that the scene is planar. I mean if you have tried it right, if there is too many depth variations within the scene it will struggle because it is meant for actually the underlying assumption of that algorithm that does the stitching is that the scene is planar okay and therefore it works if you are farther and farther off that is why right if you take a lance I mean if you take a, if you kind of go to the beach and if you take right it will come out very well because in such a wide expanse, you try doing it inside a room you know you will struggle right, the algorithm will struggle.

So the idea is that, so the geometric transformations, so the applications right as to for example right you know you can have situations of various kinds, for example you may have a situation where somebody gives you two images already taken and then what do you do with this geometric transformation, it means that right you want to understand how one can be obtained from the other because they are basically images of the same scene but that knowledge helps you to actually reverse the mapping that has happened so that you can align them right. So for example think of one image that was captured like this and then let us say the another image which was captured with actually different orientation of the same scene. Now you cannot compare them directly right because they are not aligned. So what you can do is if you know what transformation this image has undergone with respect to I1 then you can actually invert that whatever that has happened you can undo the transformation so that the two become aligned and once they are aligned right you can ask things like what has changed, for example a classic example would be like IIT Madras 40 years ago if you had a picture of IIT Madras taken a really or something and then if you want to know what how much of green cover has come in now or has it reduced, has it become better whatever it is right then you need a change detection right, you want to see what is the change and unless you align you cannot find the changes and by alignment right you want to be able to sufficiently you want to be you want to have the liberty to be able to align you cannot say that oh take exactly from that same position who knows from 30 years ago from where they took right you cannot go back to that exact same position and also you will take from somewhere around that you know roughly this was a region that was that was captured and you will be somewhere around there right and your viewpoint could be a little off from the original but you still want to be able to do the matching.

Matching in the sense that you want to be able to tell right I mean what is the change that could be that could be one application right which you can think about that is a change this one detection change change this one detection of course there are other issues here which I am not mentioning okay but let us just you know take it take it at a sort of a slightly higher level right and then a panorama right which you have all seen. So so there it is not a question of one image you could take multiple images you could take like pan it is called you could pan your camera take you know as many images as you want and then and then you create a wide field of you right just like your eyes eyes have like 100 and I don't know 170 degrees or something that is the field of you of the eye whereas your camera and all will hardly be like 40, 50 degrees. So what we see right I mean that is also the reason right why why when why when you travel right I don't know how many of you have done this so when you travel and you see a beautiful scenery right one of the stupidest things that we that we do is you know take the camera and start shooting because we want to show it to somebody else but then we actually miss the fact that what we can see with our own eyes will never be captured by your camera. So so so that way right so this panorama model is still whatever right whatever you can do with that with that with that ability that you have in that camera and that again allows you to sort of you know create a wider field of you so that you know when you everybody likes it. Then I mean you know since we are also doing deep networks right you can even think of what is called a data augmentation.

See for example you know typically right in what is called self-supervised learning right this is often used. It is also used in other places but but but one of the main places where this data augmentation comes in handy is when you want to do something in a self-supervised way. What that means is you know suppose I give you give you I mean you know see for example you know it is like saying that it is like saying that you know suppose you have very few samples. So I give you let us say you know a depth map you know with respect to right with respect to certain kind of you know viewpoint then you can then you can actually augment that right. So you can say that what if I were to translate my image right how would that is it a map look like if I rotate it how would it look like.

So all those are all those are all those are augmentations right. So so so then you might want to see create additional pairs which come from the original data but these are pairs that you are synthesizing really right. So you will have to rotate the depth map whatever you will have to do corresponding things on the pair. But the idea is that you can actually augment your data set and there you are doing something which is synthetic really right. It is not like you take the camera and rotate really because that data you do not have.

You have only let us say some from a from a fixed point of view and then you can actually do do some rotations do some translations right things that you can apply on the original image for which of course you should have the you know whatever whatever you should have the knowledge as to what will happen to the other quantity that that is of interest to you. In this case it could be a depth map right then you should know what how the depth map changes right then then you can actually augment and then so on right. So you can and then typically your cameras have what is called a soft zoom versus an optical zoom. So you have already seen that an optical zoom will always cost more because it is actually I mean it has a zoom lens whereas if you say that you know I am doing soft zoom as somebody claims 100x soft zoom right it is actually he is trying to fool you because what it means is that you know he is just trying to try to do some sort of an interpolation right because all that he has is that one image that he has captured with the highest resolution of whatever that lens or lens and the camera setup offers and after that any other thing that you blow up there is no extra information that is coming in right you do not have that right. I mean see for example if you want to see something in final detail either I mean you should go closer to that object or you should have a zoom lens.

If you just capture from here and then you know I take an image and then suppose I say that I want to see additional details how do you get from that one image additional details right but then people do that right you would have seen that you know put a soft zoom and then it gives you like an 8x and then right I mean you know depending upon what interpolation goes in people sometimes are happy with that but then that if you have actually I know right if you have been in this class then I guess you have to be a little more alert when you see those images right. If you are not in this class I would not care then maybe I just feel happy about what you saw but then right if you have attended this course then maybe you should go back and think about what they did right actually to show you that image. So this geometric transformations one thing so all this is assuming that the scene is planar there is also a special case where you can actually see the other point is when can you relate right it is not always true that you can relate images like that okay let us also understand that right it is not true that every time somebody gives you two images right you can apply some mapping on one in order to arrive at the other okay only for only for planar scenes you can do if you have a 3D scene as a special case you can do okay. Now if you have a 3D scene then the camera motion right has got to be going to say restricted I mean then otherwise you cannot do okay so yeah so I think you know those are things that maybe right I will talk about when we do this 2D geometric transformations and then there is a hierarchy right in the sense that going from simple to complex okay what kind of transits so the most general transformation that can have is what is called a 6D motion right so you have a camera right and then when you actually move I mean you can move like you know Tx, Ty, Tz right along the x, y, z axis and then you can do an Rx, Ry, Rz right I mean so which basically means that you have 6D motion right that is the most general motion that you can have so when you take two views and the idea is that you can still relate them you can translate the camera see translation is actually a big deal by the way okay in a camera okay now that we are entering the phase of this geometry right so I think you know we should start so some of these things should start sinking in now so translation is actually a big deal okay why is translation a big deal just as our eyes are there right there is a translation right and so I think of this as one camera this is another camera you do not have one eye and then right you kind of say rotate and then see the other view do you? You don't right so it is like two eyes right and there is a translation so that is the translation that we do but then our eyes do not do a homography typically why do we have this translation because we want to actually right pick up the say depth of the scene right because that is called what is called you know a parallax effect so the parallax actually means that something in the front will move faster so you can see that you know you just keep one eye closed and then right then you open the other you see how the finger shifts right which means that something which is close by will shift a lot more right and that is what is actually a disparity which we will see right when we do actually geometry and there is a formal way to actually do it and so on but then the reason why I actually brought it up is because it is not always true that this kind of a geometric transformation you cannot always align two measures simply by using one sort of a global operation okay you can align I mean even in the most complex case but then that requires more complex operations not right through this thing which I will be actually talking about okay what are called geometric transformations. The other one then we will move on to so here right no notion of depth okay even though it is geometry it is a geometry of sort of you know a different kind and then we will go to actually what is called single view multi view no single view and then kind of move on to actually stereo.

So this is the single view simply to understand how an image is formed and then the real idea is an actually solving for what is called a stereo right what that actually means is that you know so for example okay right you have an image which you take from here and then if you have another if you take the same camera or if you have two cameras right each one translated with respect to the other and in the simplest of cases right it could be so it could be simple translation along the x axis you could do other things also but actual stereo is like that I mean so it should be like on or like you know in plane so you are not supposed to tilt or anything right you are supposed to move like this you can go like this but then normally right normally the kind of this one the motion is along x and then just as the eyes are right and then right you can actually you can then start asking what is the depth and for depth right you will again kind of see come back to these features because this feature that we used in geometric transformations that is to align images but now when you are when you are talking about depth right you have actually a 3D scene now right and you are asking for example where is where is where is you know each point in the scene right and for that again right again the again the way to actually look at it is you have the camera right so for example you have a you have a point P in the scene suppose you take one point in the scene right and you are actually seeing it through a camera whose optical center is here of course typically the image plane is actually right to the back of the camera center but you can always think of think of it as being in the front so that you get a you get a you don't get an inverted image and then the same point right you also observe from let's say another camera it could be the same camera that is that has been shifted or it's just another camera and then you got like C1 and then C2 these are the camera centers right then what then what you can do is you know so for example if actually if I if I pick up a point here right in this image plane in the kind of left image if I actually pick up a point then then if I know if I know where my camera is camera center is then I can actually take this ray and and and then and then right when I can take this backwards I know that this point P ought to lie somewhere on this ray but then with this one image I cannot find out where it is so then if you can actually give me a give me a correspondence that this point lies there right in the other image again this feature correspondence that we talked about for this long right we spent so much time understanding this feature correspondence that that again right is something that we need and if you can tell me where where the actual feature point is right then then basically then you know what I can do is I can actually I can actually take this ray from C2 take it backwards backwards and then see where these two two rays intersect and and then that is called as a triangulation and then I know that right P actually sits there but this is but there's a more far no but this is very formal okay the way to do it is what is called you have to bring in the notion of a fundamental matrix there are epipolar constraints are all pretty nicely structured okay so just to give an intuition the feature point correspondence that we saw again will have a role okay 2D geometric transformations again we need them in order to be able to apply what transformation or to even figure out what transformation is exists between the two images you will again need them for stereo okay so for example so for stereo right where you want to be able to compute you know a depth map then third will be what we call structure from motion okay SFM it's called it's called structure from motion and the structure from motion is you know is more sort of you know it gives you even more say liberty in the sense that think of a camera right which you take around okay right around an object you can capture a picture from here then you go there capture a picture they go there capture a picture here from there whatever right and then you don't know you don't know where you are so in stereo typically it's like a rig okay which you actually build so you know you know that you know that this guy is exactly 5 centimeters away and so on so I mean there is a way to actually you know do the calibration but when you look at structure from motion it's totally kind of uncalibrated in the sense that what I mean is you don't know there is a difference when somebody says calibration and calibration but then take it a little loosely here that what it simply means is that I just I just randomly take a camera and move right I see an object of interest I just kind of take it from here I take multiple pictures I go around and take pictures now I don't know exactly where I was right I don't know at what angle right I actually you know no second image at what angle I actually took it from where I took it I do not know so the structure from motion actually means that you still want to be able to build a 3D model of that object that you are seeing what's called actually a point cloud so point cloud is like something that you would have seen how many would rotate and all that have you guys seen that I mean sometimes you take a point cloud and then you can rotate it and so on right so that point cloud is something which you can build out of a Lidar if you have a laser kind of you know equipment but that's very expensive right so you can actually do it with simple images I mean that's the reason why these Lidars even though they are out there they are still restricted their use is restricted to let's say industries you know the automotive companies and all they are the ones that are still using you don't see right people walking around with actually a Lidar and it's pretty expensive that is with images you can still build a build a depth map that still looks pretty good so the structure from motion is actually about computing the camera motion as well as the as well as the 3D scene so so both are unknown so you don't know where you are but then right you just know that know that I have seen seen seen the scene from here from there from there and so on and again if you can if you can get these feature correspondences right that is what will help you because right that's what will tell you that this point that you saw from this view is also that point is the same point in the second view and then it exists like that in the AC third view and so on and that actually helps you helps you understand where you are right and then once you know where you are okay then you can then you can actually with respect to that view you can see what kind of a depth map depth map you have right I mean depth map is always it right dependent on the view I change the view then the way my depth map appears will change so there it's like from this view if I observe what kind of depth map do I see then I can ask from this view right what kind of a depth map do I see and so on right and you can actually merge depth maps in order to get actually point that that is like the let us like you know like the eventual goal but in this right we will see I mean how far we can go but my idea was to actually at least you know talk about structure from motion singularly at least the math right that is involved it's very interesting and it also means that there are certain ambiguities which come up it's not true that the way you see with your eyes right you can't see the depth map like that there will be certain ambiguities in this depth map what I mean is for example you may get only a scaled version it doesn't mean that it's it's to absolute scale it's simply because certain certain things are not known how to say that so for example translation especially this is actually you know is actually a tricky thing see for example because all that you have you have are had these images right you don't have any other information if you had information about how you moved for example if you had an inertial sensor or something on the camera right which would give you by how much you moved and so if it gives you absolute movement in terms of millimeters and centimeters then you can incorporate that there are papers that also do that that actually take these raw raw data right that comes from your camera integrated with the image but but then right now right we are not even talking about all that we are not talking about gyros and inertial sense that the accelerometers and all they're just talking about images right so with an image right I mean you know if you see if you see that that a point is shifted right by by so much okay it does not it does not mean that you automatically have a sense for sense for how far away is the scene right because what can happen is the point right could have been could have been very close the scene pointed could have been could have been very close and my own my own motion is small or it could be that the scene point is very far and I I actually actually move a lot in either case where you can actually get the same shift do you do you see that see for example the parallax effect that that we have right I am saying the parallax effect just because right between one image and and the second image let us say right you find that that this guy it is at is at some location and you know that in the other image the same point appears shifted right by by some amount let us say it is a 5 pixel shift the 5 pixel shift does not mean that I automatically know where the where the scene point is okay where the actually see depth of the depth of that particular point is okay because if I if I do not know by how much I if I know the camera motion then I can do that but if I do not know it then then right you could have a situation where the where the depth where the point was actually very close to the camera okay in which case all that I had to do was move a move a little my camera I just move a little and therefore I will get that 5 pixel shift or it could be that the that the point is so very far away but then I know but then I know I need to move a lot okay so unless I know by how much I moved if that absolute number is not there by how much I moved I cannot tell whether the way is far or near right so such such ambiguities exist right so so it is not like whatever you see right now you can immediately start you know putting that to scale and but the nice thing is it gives you gives you a good feel for what you are seeing then on top of that right if you had if you had other information that can actually come in then those ambiguities one by one that you can start to get to see resolve them okay all right so I think right with that as the outline okay of this is the geometry right before we go to the next topic right so the geometry okay is is is is something that that we want to go through first.