Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-68

So, we will start with what is called structure from motion, this is just a continuation of what we have done till now just that it actually escalates it to, escalates it even higher okay, this is called structure from motion. Okay, it is a very, it is a kind of an intriguing topic and it is nice okay in several ways, it is called SFM in short and the structure refers to the 3D scene okay which you are imaging and motion refers to the camera motion. The point is that until now what we saw was something like a stereo right where we said that I mean if you had a stereo pair then you know if you kind of knew the baseline you can solve for the absolute depth. Now kind of what we are asking is what if you have to actually take a camera and sort of say go around like for example you may have let us say a fort or something right which you want to watch, which you want to sort of imagine then you want to build a 3D structure of this fort. Now you could take a camera go around right you know and take images. So for example this is like a camera you know in let us say the post 1 then this is in post 2, post 3, post 4 whatever right post n right, n camera poses you have and normally it is the same camera so you just kind of say take it around it is not limited by the way right you could also have you could also think about doing this for the case when let us say right different people okay it is a very famous monument and you have got different different images taken by different people and you want to still integrate them all together right that is even more difficult.

But let us just get a look at the situation where you take a camera you go around right this is the simplest thing that you probably do right you have to go around an object of interest it could be indoor it could be outdoor whatever it is right and you capture n views as it is called right because there are because there are there are n poses of the camera. So you have like n views of the scene right and in this case in this case right it could also happen that you do not know exactly where you are because you are just taking the camera and going around right you are not measuring that I did so much rotation I did so much translation physically and all you have no idea you just go around okay unlike this 3D okay where we said that you know it is exactly like somebody gives you a very sort of you know a constructed situation where the 2 cameras are exactly apart and then you know everything is perfect and then the z you can estimate there you can find in terms of absolute values. But now right now as you can see it automatically means that there is going to be some ambiguity now right because you do not know the exact motion of the camera that becomes

an unknown now and the camera could also involve say rotations and translations both because there is no guarantee that you are going to be able to maintain the maintain it rotation free. So with respect to the first camera right you could have a pose right the pose change in second pose change in third and so on right so you have lots of pose changes and the idea is that is that in such a case what do you do then and the simplest of this is actually what is called what is called 2 view stereo or you know 2 view is the SFM where you know see typically that when we say that when we say multi view geometry what we mean is the number of views is actually more than 3 right greater than or equal to 3 okay.

There are at least 3 views or more okay that is when this is called multi view geometry okay and there is a lot of stuff in this okay so I am going to just focus on things that are really relevant to us but this is a huge subject in itself okay multi view geometry. The idea is that you can so in 2 view right you have got like n equal to 2 okay and by 2 view SFM what I mean is again the pose is not known let us say right you do not know what is this relation between the 2 cameras and you want to be able to able to able to sort of estimate so the structure from motion that involves solving for this 3D scene as well as the camera motion right both are not known it is like a joint problem now okay and the way right it is normally done is that you sort of get an estimate of the pose of the camera first and then once the poses are known right then you sort of do a dense reconstruction right. So normally right so I think you know right let me just write down and know there are several advantages you can also ask right I mean you know what happens if there are n views are there are there are other advantages to having n number of views there are advantages too right and some of them I am going to list here. So it can give you better robustness to noise so instead of just having 2 views or something right if you have so instead of just 2 views if you had if you have multiple views right then better say robustness to noise is one thing then you can have you can even perhaps use the additional views the additional views can be used for let us say a verification purpose like for example you might say that right from this view from these 2 views I am able to say something about a 3D point. Now I might just want to use the additional view to just verify whether it is correct additional views can be used for the way for let us say verification perhaps for so it is not always true but then yes you could use it okay it is not like always you need to do that then this allows you full 3D reconstruction full 3D reconstruction what that means is what that means is right you see from this camera view point you will have a certain depth map right like I keep telling there is something called you know a depth map right it is like you know a 2D thing on which you will see that you know each image point if you get a back project where does it hit right and that kind of gives you the z component that is the depth with respect to camera 1 but then you have a depth with respect to camera 2 which is not identical to what you see from camera 1 then you have a depth with respect to camera 3 which is kind of which is again placed somewhere else because as you are moving there is a camera 4 and so on.

So you have multiple depth maps and you can actually think of a full 3D reconstruction this is what they call as a 3D point cloud I do not know right have you played around with this kind of ply files and all where you can actually rotate an object and try to see it from the sides and this and that I do not know if you have done that but that requires multiple views with just one view right you cannot see what is on that side right if you just have one view for example right this object I mean right if it is curved and if I am seeing it from here then maybe right depending upon the field of view of the camera right you can only see so much right. So here is the object and what is that okay so right so here is the object and if it has the camera has a certain field of it then you can only see so much right whereas you might want to see what is out here and what is out here and there is a decent overlap as you go across these cameras right then you have more information of the same scene point viewed from you see multiple cameras and the idea is that can you sort of integrate all of that in order to be able to arrive at something that gives you more reasonable you know reconstruction and also what we call is full 3D reconstruction by fusing depth maps. I will just indicate later what do you mean by fusing depth maps okay our idea is not to go into the details of that but yeah then you can think of fusing multiple depth maps in order to get what is called a 3D point cloud and once you have a 3D point cloud then just like the other day we talked about view synthesis you can talk about depth from let us say wherever you want right so if you just do a projection onto that view of the camera that you can get actually depth map with respect to that view then you can go somewhere else right ask for a depth map from that view you can get that. So in a sense right depth map is kind of is basically a function of the view point from wherever you are seeing right so depth map is like view dependent and you can actually fuse multiple depth maps in order to get something like a full 3D reconstruction. Then better occlusion and better occlusion handling occlusion and visibility handling right what this means is that if something is not going to see visible from some view point perhaps you know that point reveals itself when you go to another view point right and therefore you know that such a point exists.

So from one view already may not be able to see because something in the front is blocking it that is whereas when you go and see it from here then you know there is a point there right so better occlusion and this visibility handling. So all these are there but then the key challenge is this sort of an unstructured nature right in the sense that you are sort of free to go around with a camera and this is what you probably may want to do most of the time right you do not want to carry a rig with you and try to find out how much I am moving and all that right you just want to take a camera you take your cell phone you go around all right. So we will start with actually 2 view SFM because 2 view SFM is right is good to kind of start with something simple and then we will see that when you have when you have multiple views then how can you do something like a batch processing

right because idea should be to sort of combine all of them together so that one supports the other but we will start with what is called 2 view SFM.