Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-77

Then, the next thing right that I just wanted to do was this deep nets for stereo and you see SFM. So, see, so the point is right you can ask I mean, see this is the geometry part right we did because it was important to know as to how these things really work. Suppose I told you that you second as I know a deep network, if there is a disparity somebody gives you a ground truth sort of a disparity that means somebody creates a data set right where they kind of give you a disparity right you might say that you know I might just use kind of a deep network and then whatever right when if I have a stereo pair I will kind of give a stereo pair and ask the network to find a know disparity such that at the output right it actually matches a ground truth right. So, that would be that would be the way a deep network would sort of right take this problem forward, but then that is not interesting right I mean without underlying understanding what is underlying geometry without understanding the basics of what is this whole geometry business and how things how things are right interconnected and all it without that there is no point doing directly jumping into a deep network okay. So, which is the reason why we why I know as in this course right what we have been doing is we have been initially looking at what is the kind of classical theory right behind a particular thing and then right come over to deep networks because that is that is the current sort of in thing right I mean so the idea is that whatever you have learned there using that you may be able to understand things better here right so that is the goal. So, if you look at it right it is not even that old right when of course by deep learning standards this could still be considered old because every month something or the other keeps happening, but if you think about it right this is like 2015, 2016 right this is I think one was a conference paper the other is the same authors they just extended to extended the work to a journal and the idea is not to go through the right you know details of what the architecture of each network is and so on because if you do that right then you for one network right we will need one class okay that is not the point.

The goal I thought I thought right what we will do is to is to just highlight right what things people have done and right what kind of what kind of approach do they take how do they kind of look at this problem right when they try to use it using you know solve it using a deep network. So, this so this original paper right this is like this is like among the first deep learning papers that if you will so this is like stereo matching by training a CNN to compare image patches. So, also as I said before right comparing image patches right is not an easy thing I mean you can see the earlier papers if you look at it right they were actually focusing so they had so they had like you know it was not an you see end to end sort of a network right it was not like you push a stereo pair and then you ask for sort of right this one a disparity right that would be like end to end solution right I mean just give an image pair and then you hope that the output of the network I mean you will get simply a disparity out. But but it is not easy right I mean it is not it is not like well I mean I can simply come up with some network right that can do that I mean there has to be a lot of ingredients that I have to go into what this architecture should be and how I should build it it is not straight forward right but but yeah mentally right if you kind of if you if you really think about it that is probably an end to end solution where you have something in between it is not clear what should be the cost and all that.

But one cost could be that first of all right it could be that we want to do it in a supervised way that was the initial approach right because people were not even thinking about unsupervised and all right they said if you had the one of the one of the one of the key issues right with respect to using deep nets deep networks and even today is that is that right the ground truth is not there are not there is not really right that much of a ground truth available and and especially the ground truth is available that is also under very sort of what you call in a settings that are for example a kitty is okay I mean a kitty is one data set that that has a ground truth there is an outdoor scenes and there are and there are few more. But there is there is nothing that is very exhaustive and so on and you know that to train a deep network right what you need is something which is which is very very exhaustive I mean if you want to if you want to make this a network simple simpler and you want this whole thing to work well across different situations and so on then really what you need is a comprehensive data set right. Now you know people have even even done what is called I think it is called synth data set or something which is like everything is synthesized out there. But then the problem with I mean in the sense that right they can actually create situations that you want for example if you say that in the other image right I want a certain lighting to come in I want a shadow effect to come in I want you know a color change to come in or whatever right I mean I want some noise to be there right. So, all this you can synthetically add because you know going outside and trying to do it right in a real situation may not be easy right because I mean you cannot say that you know I mean you cannot create situations like that in a sort of a physical world it is not so easy right.

Whereas synthetically yes I mean you know if you if you just if you just know how to kind of render a scene right that is like the forward problem right you kind of imagine that I will have a light source here this is my 3D structure if I kind of forward project it what will I what will I see as my image and how should I take the light sources and all into account in order to get what I want. Now those kind of things have also been done, but then the problem is that becomes where so that has problems in the sense that because it is a synthetic data set the moment you go out right to an outside world the way it generalizes can have problems. So I mean it is not like it is not like people have not made headway and all right that is not that is not what I am trying to say what I am trying to say is that there is still there is still a lot more to go right in the sense that even today you cannot simply take what has been trained on one and simply use it on another situation, but for but for this kind of traffic systems and all right where you have this what is that you know automatic driver assistance systems and all there I think you know some data sets are there which are kind of very good and even for Indian conditions I think where people have gathered data sets on the road believing that this kind of an ada has right will come up at one time or the other and there yes right I mean some sort of some sort of headway has been made and then right people are actually using those data sets in order to understand and especially right people actually have captured stereo pairs and so on okay. And sometimes right the ground truth comes from comes from another sensor right I mean you know you could of course you could have a classical stereo matching algorithm which you can which you can choose right to your advantage let it take time it does not matter but then right as long as it does a good job you can get that depth map and you can claim that that is a ground truth right that is one way. The other way is actually have you know a different sensor for example you know a Lidar or something that is also looking at the scene simultaneously and then of course the only thing that you have to do is you have to mutually you can say register them because they do not have the same right they cannot be at the same point right so there is going to be so one can be next to it or whatever it or more in order typically it is kind of kept above the camera and what you have to do is you have to whatever is the 3D point cloud that is Lidar gives right that you have to kind of say register so that so that you know that from the stereo view point at how this how this scene looks like.

So that is again another thing that people have done but the problem with that is that Lidar does not give you a dense depth map I mean as you know that Lidar still has limitations 1 it is very expensive and 2 right it does not give you kind of you know a dense reconstruction and therefore right people still are not very happy so you can use it in a semi-supervised learning setup where you know some ground truth data is available which can come from a Lidar but it is not like you know everything is available okay. So all these approaches if you look at it they have all looked at it from these angles you know should I should I kind of look at it as an end to end problem or should I just look at parts of the problem for example this paper is looking at one part of the problem so like saying that how best can I tell whether a patch matches with another patch or not. So this so this network's job is only that so for example let us say so if you give this network it just takes as input 2 patches and it tries to compare them and it compares them in the manner that you would expect you know deep network to do in the sense that because you have a ground truth right you can actually you know whether these 2 are actually matching or not and therefore inherently right you are saying that even if there is some variations and all in this it is it is sort of it is sort of picking up things that it should consider as noise and sort of ignore for example if there is only illumination change maybe just things of that as extraneous noise and sort of knocks it off and says that right it is still a match whereas when it is not actually from the same same same area it is from a different area then it knows that right because there is a ground truth. So the ground truth would be like you know like a disparity so you know that you know that if I pick a patch here you know you know where which patch is the match for it I know that but then whether when you train it right you also you also train with train with other patches around it in this network especially. So if you see right what they are doing is so for example you have these 2 images and all these will actually will assume rectified image pair that means the search is going along right only search is expected only along a row and this is again solving a sub problem that means it is not end to end right.

So what it will tell is a kind of a kind of a binary sort of rate I do not know a decision where it where it is where it is sort of saying that one minute so where it is sort of saying that so where it is kind of saying that whether it is a bad whether it is a good match or not okay. And what it is doing is I think you know I believe that that that right here you have a convolutional layers then after that right these these become fully connected then there is a concatenation then after that right so in fact L2 to L8 L8 is completely fully connected and the kind of convolutional layers are coming in here this is this is again right one of the original papers and the idea was that right you compare you see you take you take 2 patches at a time and you sort of and you sort of right compare them you know and you sort of learn features right the whole idea is that you actually learn the entire bunch of weights right that you want in order to be able to able to able to arrive at a at a sort of right a decision to say whether it is a kind of a good match or not and you know the ground truth okay. The unsupervised case I will come to later so you assume that the ground truth is given so ground truth as I said right somebody gives you a disparity map so you know where this patch should match to what it should match to okay so I think right this was one of the earlier papers and there are further details and all right just leave it to you to sort of go through the paper if you are interested right to know the exact details and then this was another paper that that is from ISIP 2016 and here the idea is that right in order to increase the increase the increase the receptive field so the idea is that right the more you see I mean right you should not go kind of too local so these are ideas that right things around for example we always believe that if there is context information that you can do do a lot better rather than just just focusing on focusing locally right that is one of the reasons why you have a deep network and then we say that deeper layers enhance the receptive field right so the guy the neurons coming towards the end are actually seeing a lot more of the image than the ones right that are ahead of them. Now one other way to increase increase a receptive field in fact you will also see further works where this kind of a receptive field right becomes the what is called wide context learning so the context right so you need to so you want this network to kind of look at as much part of the image right as it can but there are but there are ways to do it right if you just increase the depth then then your unknowns and all will go up these computations go up therefore right people come up with ways by which you can actually simplify the goal kind of reach the goal but then using some simplified means and here right it is like this so you actually actually down sample the images through a set of scales right and the scales is up to you how many scales you want to use but I think in the paper they use like you know 3 scales or something so you can have different different scales and because the patch size remains the same right so again right again again it is like it is like a patch right which are which are actually trying to pick from the left and the right right images right so these are again rectified and so on and the and the and the and the right idea is to kind of use use a convolutional sub network which is kind of right which is out here so after you get this TLR right so this TLR is here so after this this is the convolutional network that follows that you have a you have a bunch of convolutional layers followed by some kind of you know kind of a one dimensional array which is like this L4 which is here right and and this this L4 coming from each one of these networks is actually concatenated and the idea is that because of this down sampling and because of the fact that the patch size is still remaining the same so it means that all the down sampled images the the the the portion of the scene that the patch is seeing is a lot more because you have down sampled the image and further if you go for even furthermore down sampling factors then keeping the patch size same will mean that you will see even more of that more of that scene right and that is the idea so you kind of use you want to use bring in the context and and right this is all coming in because traditionally right it is not been very clear as to how you should handle occlusions how you should handle noise how you should handle illumination so deep networks are kind of paved the way for that in the sense that you just know the ground truth and you say learn whatever it takes right because what it takes we do not know to kind of to to handle this in every situation so the idea is that learn whatever it takes you want to you want to get the context into account in order to be able to do a kind of you know a better job use a context right but how to use a context and you know which way do you incorporate incorporate the context right that is where one one paper to another right things change then this is this is like unsupervised okay yeah so see the others that I talked about were all supervised in the sense that somebody gave you gave you the gave you the information okay we have time so maybe right I will even write down some of these points because right instead of having to read the whole paper right you can just keep a note of what these things are so just in brief right I will just write some of these things in brief so this so only the key points right I will write okay so this 2015-16 this first paper is this I mean so you should just relate it to those papers okay so 2016 so this one is is among the first DL papers right and so input is actually is actually a left right pair okay it is a left right pair and rectified rectified and output is actually two nodes okay so which are supposed to tell good or bad match patch match right is called like patch match right so good or bad match and training is supervised training is supervised so that is you know you know the patch in the right image that should match with the left that should match with the left so you also you know this so only for such patches right you will ask for an output to be one and for the other patches like you will ask for an output to be zero that should there is a patch in the right and that should you know that patch in the right wise action match with the patch in the the left match with the patch in the left image okay and has a combination of convolution It is a combination of convolutional and fully connected layers, let us see layers, okay. Yeah, something like this I think, right, is enough and results shown on the kitty data set.

I do not have full information for everything, but wherever I have, right, I will write it down, but most of them will be a kitty or you know, what is called a middlebury data set. These were, these are usually the ones, okay. Then the next one is the one that I just now mentioned. So, there is two which is this multi-scale thing and this is to, the aim is to, the goal is to actually get the context into the, get the context into the or contextual information into the matching process. Of course, you know, I see now, right, now that you have done deep networks, you have also done a classical theory, read any of these papers, you should be able to read, okay.

That is the reason why there is no point me going through every paper and all. So, I am just here to give you a road map of what has happened, but really read any of these papers, now you should be able to pick up and read. You will be able to perfectly understand unless of course, no, because it is a deep learning paper sometimes, right, some of the architecture and all, maybe you know, there could still be some trouble, but as far as the theory is concerned and all, right, you should have no problems at all. Get the context into the matching process through a series of down sampling, through a series of down sampling or down sampled images of the, of the say left, right rectified pair and this is, so this network I think also has some, has convolutional as well as fully connected layers, has convolutional and SC layers and so on, okay. Now, let us come to the next one, which is, what is that, right, this is over, right, where we were.

Now, this is unsupervised learning, right and what is the goal here, yeah. So, here, right, so the idea is that unlike the earlier one, right, where we were taking patches, this is more or less like an end to end trainable network, okay and you might wonder, right, where does this thing about unsupervised come in, right, because you are now sort of saying that I do not even, I mean, of course, you may not do as well when you have a supervised sort of a setup, but then even to attempt something, right, at that time in 2017, right, which could, which is unsupervised, this is actually interesting. So, the, so the, so the right key idea is what is called, what is called a consistency, left right consistency. So, what is, what is really means is that, means is that, right, if you have, if you have, you know, a disparity, right, map at the output, right, which is, which is, which is what should come here, I mean, so the output has actually a disparity of this, of this particular, this one, network. Now, how do you check that, right, I mean, you do not have a, you do not have, you do not have a ground truth, right, because if somebody gave you the ground truth, you know, that would, that would, you know, that would relate those two images through a disparity map, then it is easy, right, then you can do it in a completely supervised setup, but then because it is unsupervised, right, so, so what really happens is, so what it means is that, if you, if you have, if you have a disparity map, which you have estimated, but then, right, then you want to know, you want to know whether it is correct or not, then right, one way to, one way to do it is, do is, what is called, what is called a left right consistency check.

That means you can actually warp, warp, right, one of the images and then, and then, kind of, say, try to see how closely, right, it would match the other. So, at those points where you have a disparity that is correct, right, it will actually match and at those points where the, say, disparity is wrong, right, you would not be able to match, right, this, this you are able to do because, because you have a kind of, you know, a two view situation and this kind of a left right consistency is also used when you have like a, right, you know, a monocular, a monocular situation, right, a monocular is like, you can see, this is like stereo, right, which actually means that there is a rectified pair but then think of a situation where you have just one camera, right, and you are moving, right, and you want to, that is like a single camera, right. So, monocular means a single camera and you can again ask, right, in, in such a situation can I kind of, can I arrive at a kind of a depth map, right, and there again you have to use, use two views because, because there is no other way, if you want to do it in an unsupervised way and there especially it will be unsupervised because you are just taking a camera and walking, right, and you are capturing pictures and you want to know what is the, what is the scene in front of me and typically, right, you use that, you know, using let us say, right, video sequences and so on and there again you can ask the same question and these are all typically unsupervised approaches and the thing that they bank on is this left right consistency. So, here it is simply a matter of warping but there, right, you will have to involve R and T and all that, you will have to involve the poses and all that, here it is not that complicated because this is, this is simply a stereo pair. So, so the key take, take away is this, this is kind of a confidence map, right, that you, that you get and this confidence map, right, is going to be bright in those portions where the, where the, where the matches actually turned out to be correct.

So, the idea is that, right, how do you kind of see train this network, right, end to end, manipulate the weights such that this region of, this, this confidence map, right, the, the regions that are white, right, which means that those are the places where the matches are the best, how do you actually, right, improve upon that, how do you keep. So, with every iteration you are hoping that this confidence map, right, will become, will become more and more accurate over the, over the, over the entire set of pixels, right, that, that is. So, the key take away, right, of this, of this unsupervised learning is one is its end to end, right, until unlike the earlier one, the earlier ones and all, the block matching and all will happen, the patch matching will happen but then after that, you know, a traditional guy will enter, I mean after that, right, you will have a traditional way to then compute, you know, disparity, then handling occlusions, handling everything you will have, kind of, you know, a traditional thing. So, what is called a mix of traditional and, and kind of, and see deep networks but this is completely end to end, right, there is nowhere anything coming in, so you just, just put everything end to end.