Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-85

So, the sort of a traditional algorithm right, which is for segmentation, one of the last ones that I wanted to talk about is this mean shift method. And unlike the k-means and the GMM right, where we had certain sort of these constraints, the sense that GMM right, we said that it could probably model is a elliptical clusters. k-means of course, you know has this constraint that it has to be circular and so on. Whereas if you notice here, this is a very sort of and it is elegant way of doing and a reasonably robust method, then you can see that it has been able to do the segmentation of this landscape. And you know, you are able to see that right, I mean all these things right that have been segmented as one segment, these colors and then the water as one color and so on. And right and you know, you do not see any particular shape kind of thing here right.

So, this whole segmented has actually come out as one segment and so on. Therefore, this is actually a versatile technique okay for you know for doing segmentation. And what it does is right, so in a way I mean I will kind of I will work through the math a little later, but just want you to understand right what is it. So, mean shift right, so which basically means that you know it is trying to shift the mean every time, it is a kind of right iterative process.

And it is like a gradient ascent algorithm and we will see right why it is so and the way it works is that right it actually you know it seeks modes or the maxima right in a sense or the local maxima of a PDF in the feature space. So, in a sense right what it is kind of looking at is for example, here is a feature space, it is the LUV which is a color space. So, if you had this image on the left and if you tried plotting the values of the LUV right where those points are, so you can see that you know there is clearly a cluster here, there is clearly probably a cluster here, there is a cluster there, there is a cluster there when you are going to visually look at it right. You can see that you know there are these groups that are actually existing there and the way this mean shift algorithm works is by actually modeling each one of these as actually a PDF whose mode is what you seek. Because right it is like saying that it is like saying that right no you have a group and there is actually you know a concentration.

So, where there is a maximum sort of a concentration that is your maxima that is a PDF maxima and around at local concentration you have a group. And similarly you go here there is a local concentration around which there is a group similarly right so on and so forth right. So, you can so the way so the way right you can think about it is it does some kind of a search. So, for example, you could start from any arbitrary point okay and it is called the kind of a basin of attraction. So, what that means is right if you start from here and then here is a search window and let us say that you started somewhere and then there is a vector right that is actually pointing out which way you should go in order to achieve a higher this one concentration okay that is the idea of this local maxima.

So, to say that right you are going to achieve a higher concentration what this means is that you are initially in blue and now it is indicating that right where you need to be heading to is this is this yellow dot right. So, as time progresses with the next iteration right it is so the blue has shifted to where the yellow was. Now the yellow is shifted further right indicating that there is a concentration the basin has a higher concentration in that region and then it goes there right. And then from there if you see also notice the step size keeps on shrinking right because as you keep approaching the mode the step size will shrink and then right eventually eventually it will go and sit there and then then right and then it will not move from there. So, the way to get a think about it is if you have a cluster and if you pick up any point right and if you try to if you try to say traverse.

So, this could be your mean shift path right your means could be shifting from there to there to there to there, but then they will all come to the same mode. So, all those points that belong to one cluster right will all kind of come together and come seeking the basin sort of the maximum of the basin. So, for example, here could be here could be another sort of a cluster right which comes and seeks it through. And so in a way right so in a way what you can think about is a kind of you know PDF right you can think of you know a terrain right. So, which so where you can think of a PDF that is actually generating those samples right.

So, if you think of a PDF here then you can think of that PDF is being responsible for those samples there and similarly you can think of a PDF here that is being responsible for those samples there and so on right. And that is the idea now how this mathematically pans out right is what we will see next ok. So, how this pans out is as follows. So, the way it works is like this see for example, right I mean if you had a PDF like that right suppose let us say suppose I gave you know a distribution and like I said right I mean you could have this and then you have one peak here one peak there and maybe another peak there and so on. So, if you wanted to model a PDF and if you had actually a discrete values right where you know that for example, when you take an image and then you plot right LUV values right I mean you have certain values that appear right and then you will typically write what will you do when the simplest way to do it is if it is a continuous PDF let us talk about a continuous PDF right that is what we are interested in then it will be like you know summation we shared m number of points right within that within that sort of a group then you will have like i equal to 1 to m and then you will typically delta of you know x - xi where this is this is a Dirac delta.

What this means is that if you have to write I mean you know write integrate you see f of x of course you know it will integrate to 1 and at the same time if you wanted to seek what is the value of value right I mean you know so what kind of area f of x has at a particular location and then you can think of it as the area under that kind of delta. Now the question is this is not good right this is not actually a good approximate this is not a good way to arrive this is still f hat of x okay not really this is still an approximation right. So, we are trying to arrive at an approximation what you really need to see for example right what this is almost saying is that you know if I had a neighboring point right and suppose I did not have values for that it is almost seems to think that right they would not occur at all okay. Now the idea is that typically right a continuous sort of a PDF will have a notion of smoothness around it right I mean if you are seeing a point that is occurring very likely that whatever is next to it will also occur with some finite probability very likely that right something else will also occur according to a finite probability. Now that is what is actually what a Parzen window does right a Parzen window technique what it does is it tries to model f of x by actually putting a bump on top of each of these values right that you have it puts a bump on top of them and this bump should be such that it should be a smooth bump.

So that I mean right think of it as a kind of you know a convolution operation right I mean you have values at some places and you want values in between and what you can think of doing is doing is putting a bump out there and everywhere right and then when all these come together and then if you want to know what is the value at some point it will be just the sort of you know a superposition of all those bumps right that actually contribute to that point. So think of the Parzen window like that and in that sense right so what we do is you know instead of modeling f of x which is actually a continuous sort of a distribution so density function in this case so i is equal to 1 to m and then we have a kernel k and then x - xi and then we have a semicolon h this is called a kernel and right this kernel typically needs to satisfy some simple things in the sense that k of x should be always greater than or equal to 0 and then integral k x dx right should be equal to 1 and so on in order for this to remain a pdf some simple things it should satisfy. But people have found that there is a certain choice of kernel some 4 or 5 right that are actually most ideal in order to use them here one of them of course is actually you know this one a Gaussian and there are a few others. Now to kind of think of this h is actually a parameter that controls the actually window size right I mean you know how much of a search area right so for example when you are sitting at a location how much far how far should be looking around you right and that is sort of a hyper parameter and right this is one way so I mean you must have seen other expansions for say f of x right and this is one or so this is called the Parzen technique. It is called the Parzen window technique and in a way you can also think about it like I said right you can think of it as a kind of rate a convolution operation where if you want if you add samples elsewhere and then you wanted to have the value of samples in between then you can sort of figure out right what would be a superposition of the contributions.

Now mean shift rate basically the mean shift algorithm is actually a non sort of parametric approach see if you looked at the look at the GMM it was actually a parametric approach right we said that it we modeled as a sum of Gaussians and so on weighted Gaussians right. Now this is completely non non decent parametric just because a kernel is actually parametric okay it does not mean that f of x becomes parametric okay. So you could have for example k as really a Gaussian kernel and that by itself does not make it really a parametric approach so f of x is a non parametric approach and the I mean good thing about this is right it is kind of it is actually it can do a generic clustering unlike your you see GMM and all right I mean you know which kind of look for right elliptical sort of you know clusters and so on. So generic in the sense that any shape is fine and it is kind of mode seeking it is a mode seeking algorithm right and mode seeking and mainly right looks at looks at your looks at your feature space right as something made up of individual f of x where a local mode right will tell you will tell you what is a kind of a cluster there. So it is mode seeking and you can actually show that it is a kind of a gradient ascent algorithm I will I will we will see that and then it is not really you know a generative model in that sense okay.

It is not really a generative model to be a generative and the other thing is that right it can actually get to the mode without I mean right in a sense that right I mean the idea is that you want to seek the mode right that is the most important thing. When I say that it is not really a generative model in the sense that we are not looking at generating samples okay unlike a GMM where probably you could have a very nice model which you can probably extend later in order to even use it as a kind of you know a generative model whereas here right we do not really we are not looking at really doing a computation of say f of x. The idea is to idea is to go and hit the modes of modes of modes of f of x and therefore right that is why we do not call it a generative model we call it more in terms of a mode seeking mode seeking approach where we are happy I mean right if you can kind of you know arrive at the mode okay. And go ahead f of x is a PDF x is actually continuous are your samples are for example in the feature space right for example xi could be a RGB color so you have like you know one color one color one color I mean what I have shown here is all those colors it could be in some space right it could be in RGB it could be in LUV so in that space right you plot all these points and then one way to get a look at look at model this PDF is to simply say that f of x is simply a summation of all these impulses but that is not correct right because x being a continuous quantity it is right it is not true that you know if it is occurring somewhere here then it would not occur elsewhere so that is the reason. Answer is 0 which is not correct right because that is the reason why you go for this kind of a local smoothing okay and smoothing in the sense that right the idea is that is that right you want to be able to able to sort of this is called a parson technique the idea is that you choose a kernel such that right when you actually think of it as a bump right which you can put on top of every peak that you have you have these xi values right think of a bump sitting on top of it and this bump will sort of die off think of another bump that is maybe sitting somewhere else that has its own bump and then all these bumps come together right and there is a superposition going on and if I want f of x anywhere in between I will just integrate I will just do a superposition and it will just kind of add all the contributions coming from anywhere it is like a convolution because the bump actually does not change the bump is the same so it is actually a convolution it is not it is more than a superposition it is actually a convolution and wherever you want right you can just add up those contributions.

So the way to kind of show it right is this so in general right so what we kind of do is that right in order to kind of pick this k right so this kernel there are various choices for the kernel one particular choice right of this kernel which kind of we will look at is something like k of x or k of u is equal to e raise to - half u this is one such one such one such choice of a kernel you can also have a polynomial kernel and so on. But this is what so for example if you think of k of norm of x - xi I mean if you take the vector case by h square right that in a sense will be like e to power - half norm x - xi square by h square right so this in a sense is really a Gaussian right. So right I mean so what to so okay now this is u actually greater than or equal to 0 okay so what this means is that one choice of u right which you can have is really this right which would then yield a Gaussian. Now what you can do is you can actually write your write your f hat of x or f of x right but I am writing f of x which really actually f hat okay this is still an approximation of x this is not exact okay so f of x right you can think of this as let us say some constant okay times let us say k of norm of x - xi square by h square okay this is your kernel and this is some i is equal to i going from 1 to m so this constant right we will also observe other things into this c okay or right now you can even keep it as 1 by m if you wish and then right we can actually put in throw in other things okay. Now when we say that it is actually mode seeking okay then it means that if I were to compute a gradient of gradient of f of x right with respect to x right so I would like to see what happens there okay right that is the part that will actually give us give us an idea into why this is called actually a mean shift right.

So what you can then do is you can write this as 1 by m summation i equal to 1 to m and then you can write this as k ' norm x - xi square by h square into let us say x - xi this is a vector okay all these are vectors this is also a vector right. Now I mean if you take a specific case right you can solve this but you do not even need that then what you can do is suppose we say suppose we indicate k ' of sorry g of x g of x is equal to just to this is simplify things k ' of x right. Suppose we just replace this by something that is more easy to that is just some function g of x then 1 by m is summation so this becomes 1 by m then you have g of norm of x - xi square by h square into x - xi right and this you can now split okay so this you can write as 1 by m where yeah yeah we are correct yeah right I mean this is going to be 2 okay. So then we will just simply write make this into some c okay just make it in some constant that will absorb everything and then what you have is okay so let us just split this as c into summation g of norm of x - xi square this by h square okay and into x - summation xi g of norm of x - xi square by h square right. So these are summed over i okay and then what we can do is we can actually divide and multiply by so we will say that is c into g of norm of x - xi square by h square right and then we will also have this term this guy right out here which is like summation x to g of norm of x - xi square by h square okay.

Now this x right I can actually take it out right because this x is not dependent on i and actually take that out whereas here I cannot take it out right this guy remains as xi g norm of x - xi the square by h square the whole thing I will divide by g of sorry this summation okay summation over i summation over i g of norm of x - xi square by h square okay. So if you so yeah so this summation is also over i so if you actually right think about it this right then think about this then what will happen is on the left right you still have gradient of f of x and on the right you have like c into summation over i g of norm of x - xi square by h square okay this is one term and then into now if you see here right this and this will actually knock each other off right and therefore you will get okay this multiplied by x so the first term will become x the second term will become - summation xi g of norm of x - xi square by h square by summation i g norm of x - xi square by h square okay right that is what it will be. Now this for most for so the kind of the kernels that we choose right this is actually a positive quantity okay this is k ' of x and the way we choose a kernel is that this is a positive quantity and what happens is right this term here right this is actually called the mean shift why do we call this the mean shift is because it is because right I mean if you kind of think about it right you are sort of looking you are sitting at some location x right so the way to kind of think about it is you are sitting at some location x and then you have a weighted mean of mean of all the points within that window which is coming from the right and this shift right and this mean shift is saying that if you were at x old right then you should be moving on to an x new which is like x old plus the mean shift right that is what in that is a graphical thing about showing right how you move. So the way you move is that you could you are at x right now and what it is saying is that you have to move from there and you have to move by an amount which is like the old x plus the mean shift amount okay that is why it is called the mean shift because you have a weighted mean on the right right right is a weighted mean right this is a weighted mean quantity and it is saying that right you need to get a shift right in order to be able to in order to achieve a maximum for this guy a gradient of say f of x and if you think about it right you can actually think about it as grade m and I can actually bring the quantity on the right on to the left and I can write this as gradient f of x by c into summation i okay g of norm of x - xi the square by h square and then I have write x - this right summation blah blah okay this is my actually mean shift quantity right. What do you see here? See when you write okay see one way is that right I mean there is actually a convergence proof at which I am not showing here there is a there is a convergence proof right which shows that as you keep doing this iteratively right so for example so the way this works is that you have like x new which is like x old plus a mean shift okay.

Now if you think about this mean shift quantity and now did you see something here this is the mean shift what do you see here? See I am writing x new as x old plus the mean shift right that is how I am trying to show that this is an ascent algorithm this is actually a gradient ascent with something special going on I am hoping that somebody will tell that what is going on? No it is like you started with started somewhere and you want to go somewhere right you want to go such that you reach the maximum of this f of x where you are seeking the mode of f of x right so think of a group of points that you have you have got right different groups of points right and when you have a group that basically means that there is a density of points there right which means that you are trying to seek the mode because around the mode is where all these things are grouped right that is the way you see it so which means that all these points belong to that group. So you are starting from somewhere right let me take a point from this group and I take a window and I am moving right so this window is this mean right this x i g this is over m number of samples right so that m number of samples is being computed over a window and I started from some x old and now it is saying that move such that x new becomes x old plus the mean shift the mean shift is the old x so x is like x old - the - is the weighted mean think of this x as x old in this quantity okay. Now what I am saying is so and then you move forward okay now what will happen is initially your step size okay that is what I did not want to use that word but let me use it now but something more than that yeah you are right think of that as a step size but something more is happening it is actually it is an adaptive step size it is an adaptive step size it automatically decreases this the quantity on the right right will keep on falling there is actually a you know a convergence proof that shows that the right the mean shift will actually you know go to 0 eventually right as you keep on iterating the mean shift will eventually become 0 that actually means that the point at where you wanted to come to the mode you have reached there and you just stay there after that you do not get any forward movement. So the way to think about it is if you think about this mean shift here right that is like you can think of this 1 by c this whole thing as your alpha the step size for the gradient right replaces here you know with x old plus some alpha times a gradient of you see f of x that is how you would do right okay so you move by actually the mean shift amount but then moving by the mean shift amount is not something that you are just doing like that it is actually a gradient ascent okay it is actually a gradient ascent which takes you from anywhere that you are any point that you take from that cluster if you move it will actually take you to the local mode there and this f hat of x that you are modeling is that local cluster right you are not trying to model that is why I said that is not a generative model this is like locally modeling each cluster you will have a mode for one cluster you have another f for another cluster another f for another cluster it is not like you are modeling an entire set of data sample right like you did in g m m g m m it was like you know you had a complete sort of a distribution right f of x like you know summation of all the Gaussian that is not the way it is done here this is like local mode seeking you think of this as 1 bump think of that as another bump think of another and then right each one you have to seek a mode and I mean the way to think about it is you have a distribution sitting there from where those samples are coming right one distribution there from where these samples are coming another sort of a distribution there from there from where those samples are coming another sort of a distribution here from where those samples are coming and you try to seek the mode of each one of these. Because once you have the mode okay then then you know that all these points belong to the and they are the nice thing is wherever you start right if these points belong to the cluster they will all come there it is like a basin of attraction they will all come head to that mode and then from another cluster they will all come and head to that mode that was that I mean.

So well some people show simulations and all I do not have an actual simulation, but that slide right it is kind of sort of showed you right what it means. M is the total number. M is the number of samples within the window, number of samples within the window of choice. Now that is the only hyper parameter here, how much you should see around you, how many samples should be here because I mean you cannot take a window that is too big right then it would not make sense because then you will encompass multiple clusters. So that is still a hyper parameter okay so that you have to choose carefully, but as long as you do that right everything else sort of moves very nicely and there is no sort of a restriction that a cluster should be this shape, that shape it can be anything.

Okay so what I would so this alpha is actually an adaptive step size, this is an adaptive step size and there is a theoretical convergence proof theoretical convergence proof that m s that the mean shift will go to 0. Yeah no no h is a hyper parameter so the window function is a function window is a function of h, h sort of tells you so h is like you know telling the influence right how much will you so this norm x - x y square if you do not have h right then it will sort of it will treat every x i in a certain way if you now increase h right then if you increasing h will mean what? Then it will mean that if something is what does it mean? So if something is close right then it will have to be really close otherwise if h is smaller then it will become e power - right, no it is just the it is opposite yeah so whatever so that way you can control right about what you what I mean how you want to treat an x i that is near to x how much weight you want to give that will that will be that will be as a dictated by h right. So it means that if you have your h large right then it will mean that even something farther off will be not equally but then yeah I mean that will also be taken into consideration whereas or else you know it will just knock it off very fast right it will be like a very local window versus a spread out window. So window is directly a function of h and depending upon the choice of the kernel you know how you the window size will take shape. So theoretical convex proof that m will go to 0 as the as you write iterations progress and this is a gradient ascent this is a variant of actually a gradient ascent algorithm I call it a variant because the step size is adaptive variant of gradient or not gradient descent gradient ascent okay.

Yeah so right I mean so I showed you some examples over it just to just to very quickly quickly go back and show those examples. So here is how it is right so you model so you have an image like this then you plot the plot the feature space and you can think of think of all those modes right which are actually seeking locally in order to be able to you know get your cluster and and right this is how it will come. So you take a point right and then if you follow the mean path right the mean shift path it will it will eventually end up in the mode here anywhere you come from the only thing is right you will have to do it for every point I mean that is what makes it computationally but there are kind of ways to ways to not do it in a brute force manner but really if you think about it every point you are sort of trying to see where it goes and then right after that right if you do if you take this image right I mean you can see that any shape is okay you can you can cluster very well these mountains right. You can see that you know these shapes you are not modeling them as elliptical or anything right any shape is okay. So that is the strength of mean shift which is why which is why I thought right I show you which is why I did not want to miss it right you know miss sort of explaining what it is and and especially right the kind of gradient ascent part right it is nice to to kind of know how it works alright.

So I think with that right we are actually done with done with all the kind of classical not all right the most the most most relevant ones I do not know what I did the most relevant ones and a deep networks right for both optical flow okay which we have already done as well as the segmentation problem right. So I am just I am just going to kind of right orally tell okay what those points are okay to the to the extent possible okay unless something really needs to be written okay.