EPISODE 351

[INTRODUCTION]

[0:00:00.8] JM: Video object segmentation allows computer vision to identify objects as they move through space in a video. The DAVIS Challenge is a contest among machine learning researchers who are working off of a shared data set of annotated videos. The organizers of The DAVIS Challenge join the show today to explain how video object segmentation models are trained and how different competitors take part in The DAVIS Challenge.

This is a very important topic and will only get more important as self-driving cars come to be an everyday sighting because if you have a self-driving car, it needs to be able to segment different objects that are coming into the video stream and identify pedestrians or other cars or anything on-the-fly. A good companion to this episode use our discussion of convolutional neural networks with Matt Zeiler, which is an episode that's linked to in the show notes.

Software Engineering Daily is looking for sponsors for Q3. If you're company has a product or service that is marketed to developers or if you're hiring, Software Engineering Daily reaches 23,000 engineers listening daily and you can send me an email, jeff@softwareengineeringdaily.com. Thanks as always for being a listener to this show.

[SPONSOR MESSAGE]

[0:01:35.9] JM: Artificial intelligence is dramatically evolving the way that our world works, and to make AI easier and faster, we need new kinds of hardware and software, which is why Intel acquired Nervana Systems and its platform for deep learning.

Intel Nervana is hiring engineers to help develop a full stack for AI from chip design to software frameworks. Go to softwareengineeringdaily.com/intel to apply for an opening on the team. To learn more about the company, check out the interviews that I've conducted with its engineers. Those are also available at softwareengineeringdaily.com/intel. Come build the future with Intel Nervana. Go to softwareengineeringdaily.com/intel to apply now.

[INTERVIEW]

[0:02:30.7] JM: Federico Perazzi, Jordi Pont-Tuset, and Sergi Caelles are organizers of The DAVIS Challenge. Guys, welcome to Software Engineering Daily.

[0:02:39.5] FP: Hey, Jeffrey. Thanks for inviting us.

[0:02:42.8] JM: It's great to have you here. Today will talk about video object segmentation and The DAVIS Challenge, which is a competition around video object segmentation. Let's start by describing the problem of video object segmentation. What does that mean?

[0:03:00.4] FP: Video object segmentation is a computer vision task that aim separate foreground objects from background regions in videos. I can tell you an example, you have a lion running in the wild, you want to take the segmentation of the lion. Meaning that you want the [inaudible 0:03:17.2], the contours of the lion in order to separate it from the background, which is probably like the Savanna, whatever, in Africa.

The idea is that, or video segmentation belongs to the bigger umbrella of like video understanding related task. You could see segmentation like video object segmentations related to video classification or object localization. It's all about extracting information from videos.

[0:03:48.2] JM: If I have a video of a lion running across the Savannah with bunch of grass and trees that are occluding the image or I've good footage from a self-driving car and there's lots of other cars around and I want to separates a pedestrian from the other images in the video feed and I want to be able to track that pedestrian specifically, these are some of the types of applications where video object segmentation might be used.

[0:04:22.5] FP: Absolutely. For the task of self-driving cars for example, you can decouple like the localization of the pedestrian with the segmentation. The localization means, okay, you teach an algorithm to find out where the pedestrian is and to basically place a bounding box around it. The bonding box is a rectangle, basically enclosing the shape of the pedestrian.

To get even more information, to get more accurate information, you want to basically classify each pixel within this bonding box whether it belongs to the street or the pedestrian, and this is where, for example, video object segmentation comes in place.

[0:04:58.6] JM: When we're trying to do object segmentation within a video, if you take a video of the Savannah and you think about all of the objects in the picture, you could break it up into blades of grass and trees and you could go infinitely granular. You could go down to the molecular level. How do you define the bounds of what is being defined as an object? What is being segment?

[0:05:24.0] FP: In this particular case, we have our own definition of object to be segmented, which is basically a moving — How would you say? It's a moving object which is with dominant motion. This is our definition of object. We respect to the foreground. This is, again — It's like a subjected definition that we decided. In the challenge, for example, we provide the input object. We tell the user which object do they have to segment, which is the lion and the grass, in general.

There is an extension we could say or something really similar to object segmentation which is called semantic segmentation, and this aims to classify each molecule which is basically a pixel of an image we've labeled. In this case, the task of semantic segmentation will be like to classify each pixel belonging to the lion and each pixel belonging to the grass, and that basically gives you full knowledge of the scene.

Yeah. Sorry, Jordi.

[0:06:32.4] JPT: From a more philosophical point of view, I would say, so semantics like lion are in a hierarchy, let's say, in a hierarchy of concepts, let's say. You can by saying that that's a Savannah so that the whole image can be represented by that. This would be the parent of everything, the parent node of that hierarchy, and then you can start breaking this up step by step and saying, "Well, in this Savanna there are the foreground and the background maybe, or the animals and the plants." Within that, you can say, "Well there are the elephants and the lion there. Within the lion, you can get in the [inaudible 0:07:10.2] and there's a head and then

there's the foot here. Within that you can iterate that until you get to the molecules as you were saying.

As Federico was saying, in our case we define which level of this hierarchy you want to go by saying this is the specific object that we want you to track because we are giving you this object on the first frame in the case of The DAVIS Challenge. In general, you will need to break the scene into this hierarchy that we were calling about. That's why sometimes segmentation is said to be hierarchical because you can break it into all these levels of concepts from the molecules to the whole scene, let's say.

[0:07:52.4] JM: In this challenge, the task is you've got, let's say, a video. In reality, you have a set of videos that you can run your own algorithm on. The general problem is you've got a video and let's say it's a thousand frames and the first frame is densely annotated which is let's say you've got a picture of skateboarding park and you've got a single skateboarder, his entire outline is colored red, so you can understand exactly what are the contours of the object that you want to track. Then in the other 999 frames after that, it is not annotated, and the job of the algorithm that everybody is writing to try to compete in this challenge is to be able to track the object that has been annotated in the first frame throughout the rest of the frames. Is that an accurate description?

[0:08:52.4] FP: Perfectly accurate. Yes.

You said that it was densely annotated in the first frame. We do a distinction here. We say that DAVIS is densely annotated because it's annotated in all the frames. If you have a 1000 frames at 20 frames per second in a video for instance, you have all those 20 frames in a second, all of them annotated, completely densely. That was one of the main differences from our work to previous approaches in which only one out of, say, 20, or one of out of 30, so only in every second frame was really annotated and then that limited a lot how the algorithms were evaluated and we're trained on these dataset because you had a holes into the annotation, let's say. We say that we are densely annotated because we annotate all the frames. That's one of the dimensions.

The other dimension that you were saying is that if at every frame that we are annotating, we have pixel level accuracy. We also differentiate from other databases in the sense that we say that they are pixel accurate in the sense they are really really low-level precise contours of the objects here, not only like a simple polygon around the object or a bounding box would be the worst case. In our case, every single pixel has been looked for by a human and has been decided whether that was foreground or background, and so you can really have a very very high precision on this sense. At every pixel at every frame, we have pixel level accuracy and we have a densely annotation in the sense that all and each and other — Every frame in the sequence is annotated.

[0:10:38.4] JM: Okay. You have described the competition, and I'll probably ask you to describe it a little bit more just so people can get acquainted it with more, because people who are hearing this for the first time. But let's step back a bit.

The DAVIS Challenge which you guys are part of is this competition for video object segmentation. Take a step back and describe the rules of this competition. What are the parameters and who enters it?

[0:11:06.7] SG: Basically, anyone is welcome to enter, either universities or companies. Anyone is welcome to enter. We have a [inaudible 0**:**11**:**16.3] lab website in which you can just simply register then you can download all the dataset annotations for the training set. Also, the images and the first mask, the annotation of the first frame of each video sequence, then you can process it offline and then you'll submit online your results and the server does reply to you with your performance on the dataset, let's say.

[0:11:45.4] JM: Okay. I think this clarifies things. If I'm a competitor, I download the dataset and the dataset has that — Like I said, like a thousand frames, and the first frame is annotated and then the rest of the frame are not. On the backend, you have all of the frames annotated so that you can measure the performance.

[0:12:05.3] FP: Exactly. Yeah.

[0:12:06.7] SG: We also give a training set which is fully annotated and a test set for which we only give one annotation in the first frame. The competitors are allowed to use the densely annotated videos during the training phase and they can only test their performance given the first frame. Of course, in the backend we have all the annotations and we can measure the performance over the entire video.

[0:12:35.3] JM: On the backend, you also have more fully annotated sets that are not disclosed at all to the competitors in case they write an algorithms that over-fits for the dataset that they were given.

[0:12:46.9] JPT: For this, we have — Let me explain exactly all the sets that we have. First of all, we have what we call the train bal set, which is in our case 50 sequences for training, 40 for validation. These 90 sequences, you have all the frames and all the annotations for all the frames. This one you can use it however you want to train your networks, to tune your algorithms and you can use if offline.

Then we have another set of 30 sequences that's what we call the test development, let's say, in which they are already available now. For these ones, you can download the whole videos, the frames, and then only the annotation of the first frame so that you cannot train on the frames that are evaluated afterwards.

We call it test development because you can submit as many times as you want your results. You can really try it and you have a number, you saw, "Well, it's good," "it's bad" you try another algorithm, you submit it, and you can submit it as many times as you want without having the annotation.

With these, as you said, people could over-fit to it, could really tune the parameters of their algorithms to really improve, improve, improve, and let's say by brute force solve the algorithm. To try to avoid these what we have is another 30 sequences. In total we go up to 150. These last 30 sequences for now, nobody has them. We have them on the backend only. Nobody has the annotations and nobody has the sequences, and they will only be released for one week during the challenge itself. That's why we call this one the test challenge.

These ones will be 30 sequences in which we will also release only the annotations of the first frame, and then people will have only five trials, so they will only be allowed to submit five times to the server, and that's the one that will be finally — The result of the challenge will be evaluated on these last 30 sequences.

Since people will only be allowed to submit five times, and since they will only have with them for one week, let's say, we believe that with these, we are preventing people over-fitting to it. If people over-fit to the test development, well, we don't care that much because then for these final competition they will only have this set five times and one week to submit.

[SPONSOR MESSAGE]

[0:15:28.9] JM: At Software Engineering Daily we need to keep our metrics reliable. If a botnet started listening to all of our episodes and we had nothing to stop it, our statistics would be corrupted. We would have no way to know whether a listen came from a bot or from a real user. That's why we use Encapsula to stop attackers and improve performance.

When a listener makes a request to play an episode of Software Engineering Daily, Encapsula checks that request before it reaches our servers and filters bot traffic preventing it from ever reaching us. Botnets and DDoS attacks are not just a threat to podcasts. They can impact your application too. Encapsula can protect your API servers and your microservices from responding to unwanted requests. To try Encapsula for yourself, go to encapsula.com/sedaily and get a month of Encapsula for free.

Encapsula's API gives you control over the security and performance of your application. Whether you have a complex microservices architecture, or a WordPress site, like Software Engineering Daily. Encapsula has a global network of over 30 datacenters that optimize routing and cacher content, the same network of datacenters that is filtering your content for attackers is operating as a CDN and speeding up your application.

To try Encapsula today, go to encapsula.com/sedaily and check it out. Thanks again, Encapsula.

[INTERVIEW CONTINUED]

[0:17:13.4] JM: As we said, every object in these videos that we're going to be tracking is annotated with an object mask and annotation — I think when people think of the word annotation they think of a text annotation, but this is actually a contour — It's like an outline or a mark. Describe what an object mask is, because I think people who are listening to this probably still might be having trouble understanding what it actually means to have an image to have part of a video be annotated.

[0:17:46.3] JPT: Yeah, annotation can mean a lot of things, as you said. What we understand in our case is, as you said, is a mask. An image is a set, is a matrix, let's say is a bi-dimensional set of pixels. It's a set of pixels. A mask for us is just dividing these set of pixels into two sets. These pixels are object. These pixels are background.

For us, an annotation, it's a mask saying, "These set of pixels is background. These set of pixels is foreground." You can see it as a binary image, black and white, in which black is background, white is foreground, and each and every pixel has a value either one or zero, either black or white, in which it says whether it's the object or it's the background. Each object is one of these masks.

Let me point out now here, one of the differences between DAVID 2016 and DAVIS 2017, which is that for DAVID 2016, as Feredico said, we had a single object with dominant motion, and so we had a single object per video.

In the 2017, we've added one of the challenges, another challenge that is that we have multiple objects per sequence. Every sequence has multiple, for instance, persons. As we're saying before, the skater, we have annotated the skate and the person. We have multiple annotations or multiple masks for every frame, for every sequence, and then the challenge is not only to track one of these objects, but multiple of these objects in a single sequence.

[0:19:24.7] JM: You're describing in the 2016 competition, it was — If you had a video of a pedestrian walking on a busy street, you could annotate that image. If you just look at every pixel and if you assign every pixel as zero or a one, all of the pixels that include the pedestrian you might mark with a one and all of the other pixels you might mark with a zero.

In the 2017 competition, on that busy street, maybe there's also a car and then you would mark the pixels that have the pedestrian with a one, the pixels of the car you would mark with a two, and then all of the other pixels you would mark with a zero if you were annotating that.

[0:20:13.3] SG: Yes. What you described is something like we assign the semantic, right? To each of the pedestrian, we assign a one because they belong to the same semantic class. While to the car, we assign two because it belongs to another class. In this challenge, to each pedestrian, we assign a different flavor. We discriminate the instance. The pedestrian is the object and there are multiple instances of this object.

You see the difference? It's a bit like between — The lines are fuzzy, but it's like something in between semantic segmentation and instance segmentation. Semantic segmentation, you associate the pixel to a class level. Instance segmentation, you associate the pixel to the number of instance. If you have five pedestrians, it's going to the be one to the first pedestrian, two to the second, et cetera.

[0:21:12.1] JM: This challenge is semi-supervised. People who are familiar with machine learning might understand why it's semi-supervised, but for people who are not familiar with machine learning, explain why that term semi-supervised applies to this problem structure.

[0:21:30.8] FP: Semi-supervised can be understood in many ways and it can be misleading in our case. In our case, we give a very specific definition of why we say it's semi-supervised, and it's because we need the input of the first frame. We say we give you the input of the mask that we want to track on the first frame. That means that there's a human interaction in there that has to annotate something in the sequence, the first frame.

When we call about in generic in machine learning, when we call supervised or unsupervised, it's whether the algorithms for learning they have these annotation, these interaction of the human, whether they take it into account or not. For example, let's say the first — The start of computer region where we were detecting edges, let's say — With a Canny Edge detector, let's say, we were just doing an edge for us or something that the intensity color from one pixel to the

other changed a lot. That's completely, we'd say, unsupervised because it's just — We define a set of rules and it's not even with machine learning, let's say.

Supervised would be in this case to take a set of annotations of what you want the result to be. In this case, edges. Then apply machine learning and learn from that. To unsupervised — When we were saying semi-supervised, it was just to separate from the unsupervised case in which you are given only a video and you expect the algorithm to take the segmentation out. That would be our unsupervised view of video segmentation.

We call semi-supervised as we are giving you the mask of the first frame, that's why it's supervises. Then we are asking you to give the rest of the annotations of the frame.

[0:23:27.9] JM: If you were to take — Let's say you're talking a semi-supervision, the first frame is supervised. It's heavily annotated. If you were to annotate two or three more frames, does that make the accuracy go up? Does that significantly make the programmer's job easier?

[0:23:48.0] SG: It depends on the algorithms because there are some algorithm would not make any use of multiple annotations. Some others that use training that train, for example, online. Giving the developer ground [inaudible 0:23:59.2] that for sure, that would like increased performance. There are a couple of example of papers that recently come out to [inaudible 0:24:06.9], which explains like the first frame annotated. I think they demonstrate that given more annotation, the performance are quickly increased.

[0:24:16.0] JM: We did a show recently about Clarify, which is an image recognition and video recognition API. We discussed some of the basics of convolutional neural networks and edge detection and then figuring out what an object and an image or a video might be after you take these edges and then you abstract to higher and higher to understand what these edges might compose. I'd like to talk about some of the algorithms that might be used in this composition, because we've been talking about the structure of the problem, like what is the way that the competition is laid out. Now, I'd like to discuss some of the approaches that competitors might take. Maybe let's start with what are the most naïve strategies for approaching the densely annotated video segmentation problem.

[0:25:14.2] SG: The most naïve structure, you take the first frame and you copy to all the other frame. That would be the most naïve strategy, which still might give you reasonable result if the object is like only moving around an anchor or something like this.

[0:25:28.8] JM: So if it's a sloth.

[0:25:30.9] SG: Exactly. That would be great. A better approach would be compute the relative motion between the two frames. If you can estimate optical flow, which is a technique for computing like a mapping between pixel of the frame times zero to time one, if you have this motion, if there are these information, you can propagate the mask forward. This also would be a naïve way to compute the segmentation while it could actually give some reasonable results. It's [inaudible 0:26:02.1].

[0:26:03.4] JM: Okay. Then what are some more sophisticated strategies?

[0:26:06.9] FP: More sophisticated strategies makes you — The most promising right now, and then I'll get Sergi and Jordi also comment on that, but the most promising strategy for sure, they use deep networks. The best results that were reached in the 2016 dataset, they were all using deep convolutional networks, deeper architecture designed for segmentation of static images. Of course, the problem on video segmentation, it's a bit more — It's not more difficult, but video gives you more information, and we want to leverage these information.

I assume that many of the competitors will try to leverage multi-frames. For example, the previous information that is given, that is estimated at the previous frame. For example, they could have convolutional neural networks, with feed-forward networks, with recurrent neural networks, like LSDM, to remember what was the mask at the previous frame and leverage this information to the next to get more accurate segmentation. I don't know what…

[0:26:57.6] SG: Yeah. Of course, the general, and I think that's not only for video segmentation but for all fields in computer vision right now. It's been invaded by these convolutional neural networks that we all talk about. That's definitely the case also in video object segmentation. I would say that there are two approaches, two main approaches I would say right now that are leading the competition, I would say, in 2016. That would be ones that take these motion

information into account and to try to propagate the information from one frame to the other. The other approach that if instead of propagating the information, just learn the appearance of one object in particular using a neural network. So learn that I am trying to follow a person with red trousers.

Then at every frame, without having the motion into account, just look for these model and look for that particular model on to the next frames. This is, I would say, the two leading approaches in 2016. We'll see what we see in 2017.

[0:28:29.1] JM: I think that I have trouble with and I think maybe a lot of the listeners have trouble with who are people who are just traditional software engineers. They've just been writing business logic, and maybe today they're trying to get into machine learning or they're at least trying to build an understanding of what goes on in machine learning.

The idea of these multi-layered networks where you have sequential processing in these different layers and you're handing off information from one layer to the next layer and then there's some sort of feedback loop where it's learning overtime how to do this better. It feels like a very foreign type of programming. Could you just give us — I'm trying to get better at having shows about deep learning, but I really don't have a great understanding of it. Can you give me an understanding of how information propagates from one layer to another? How are you representing it in code. Yeah, just describe the process of building a convolutional neural network for a software engineer who is not an expert in this.

[0:29:33.2] FP: Basically, it's quite — There are a lot of mat behind, but there are some frameworks that help you to do that so you don't have to go to really the deep level, to the core level of multiplication of matrices. For example, there is one of the first frameworks that boosts the deep learning with a café that we are still using, but then there are many companies also have prodigies and they tried to their own framework. For example, one frame that is pretty famous is TensorFlow, but mainly all these frameworks, what they try you to do is make your life easy and try to describe one of these convolutional neural networks, then lines are like 20 lines, because otherwise if you have to write by hand all the multiplications that are behind the scenes, it would be more tricky.

To do accordingly, also one important thing that boosted the deep learning community and now their performance of the algorithms was they usage of GPUs because all these multiplications of matrices, the convolutions can be parallelized, like GPUs are the perfect fit to implement this kind of algorithms.

[0:30:50.6] JPT: Let me just say that if you want to write an implantation on convolutional neural network or a neural net, let's say, this is just a bunch of matrix multiplication and transpose, right? Plus some no-linear function, like exponential function or the tangent. As Sergi was saying, of course, there is a lot of optimization and you need that to leverage the massive amount of data that is available right now. As he said, there's now a better way that implementing yourself, like TensorFlow or Caffe.

[0:31:28.7] JM: If I'm a user of TensorFlow or Caffe or whatever. Describe my workflow. What is that actually look like? For example, if I'm approaching the problem of densely annotated video segmentation, what is the day in the life of a programmer who's writing a model trying to solve that problem?

[0:31:48.6] FP: Usually, you have like a certain architecture. For example, one of the modules like — One is BGG and then you take — There are many parameters in these architecture, these neural networks. Then in order to achieve better performance and to need less data for the task that you're going to currently approach, usually what is like called pre-training. You take the network that was training for another task. For example, object detection or like image-labeling and then you use that information into your problem and you just like fine-tune. That would be adopting neural network architecture to your specific problem.

[0:32:35.2] JM: For somebody like me who is trying to get an understanding of deep learning, what would be your suggestion for the best researches? Because I know there's like an overwhelming amount of resources. Is there some particular tutorial or set of papers or book that is the best place to start?

[0:32:56.7] FP: Yeah. There are many other sources. For example, the one researcher, Andrej Karpathy that has many, for example, block bust where he tries to explain in quite an easy way and give examples of how this work. There's also, in my case, I found quite useful online, these

massive online open courses. For example, in Coursera or Udacity where there are many courses where you can actually learn.

[0:33:24.0] JM: Right. Let's get back to The DAVIS Challenge. Describe the phases of the competition.

[0:33:31.8] JPT: Right now we are on the phase of the test development that we call, so it's from now until June. In this one, people can download the current dataset. The 30 sequences of this phase, and they can submit — Publicly available, there's only the annotation for the first frame and they can submit as many times as they want their results on to the server and get the evaluation as many times as they want. That's what we call the test development phase. It's for people to tune their algorithms, to unseen sequences.

Then in June, for a week or 10 days, we will have the test challenge itself in which people will be allowed to submit their results five times. From the best of these submissions, where the winner of The DAVIS Challenge will be decides. Then in 2017, this year in Hawaii, we will have a workshop in which we will gather all the submissions and we will also have some very good keynote speakers and we will discuss about all the results that we've got, how people approached it. Discuss what worked, what didn't work, and then give the prices also. Think about what to do next and what to do for DAVIS 2018.

[0:34:59.9] JM: Who are the people that are entering these contests?

[0:35:03.2] SG: I guess researchers from universities, like Ph.D. students. There might be also like single individuals interested in the topic. Of course, our challenge doesn't give you the exposition of a competition on kaggle.com for example, this big machine learning like website.

You might also gain some visibility. We have good sponsor. We're going to have a presentation. I guess we have heterogeneous. We'll have an heterogeneous population of competitors from different backgrounds.

[0:35:40.9] JPT: I would say in general, researchers, but both from academia and from industry I would say.

[SPONSOR MESSAGE]

[0:35:53.9] JM: Spring is a season of growth and change. Have you been thinking you'd be happier at a new job? If you're dreaming about a new job and have been waiting for the right time to make a move, go to hire.com/sedaily today. Hired makes finding work enjoyable. Hired uses an algorithmic job-matching tool in combination with a talent advocate who will walk you through the process of finding a better job.

Maybe you want more flexible hours, or more money, or remote work. Maybe you work at Zillow, or Squarespace, or Postmates, or some of the other top technology companies that are desperately looking for engineers on Hired. You and your skills are in high demand. You listen to a software engineering podcast in your spare time, so you're clearly passionate about technology.

Check out hired.com/sedaily to get a special offer for Software Engineering Daily listeners. A $600 signing bonus from Hired when you find that great job that gives you the respect and the salary that you deserve as a talented engineer. I love Hired because it puts you in charge.

Go to hired.com/sedaily, and thanks to Hired for being a continued long-running sponsor of Software Engineering Daily.

[INTERVIEW CONTINUED]

[0:37:25.2] JM: It's interesting that there are some problems where the idea of competition and crowd-sourcing really make a lot of sense. I think machine learning where you have these datasets but it's very unclear what is the best strategy because there are so many different strategies that somebody could potentially take. This is a good fit for a competition, similarly a competitive programming competitions where they give you something like, "Here is a 1000 words from the dictionary and the description, or the definition of each of these works. Make the best compression and decompression algorithm specifically for dictionary entries." People take all kinds of different approaches. What are the types of problems in computer science or

software engineering that are a good fit for this crowd-sourced competition model? Does my question make sense?

[0:37:25.2] FP: Yeah. I'm not sure if we understood the question. What do you mean by crowd…

[0:38:36.9] JM: My question is. If I'm an engineer at Google, for example, and I've got a minor bug to fix in Gmail. That would not be a good problem to crowd-source to the entire company and do a competition for because there's probably not that many different approaches you could take to solving a bug. Then on the opposite extreme end, you've got something like densely-annotated video segmentation where you've got a giant dataset, you've got a million different approaches you could take. Then in the middle between those two different types of problems, you've got a huge array of different types of problem sets where there's varying degrees of subjectivity.

What I'm wondering is what classifies a problem in computer science or software engineering as being something that is approachable from the wisdom of the crowds, because that' essentially what you're doing with DAVIS, is it's a wisdom of the crowds approach to let's try to find the best way to approach this video segmentation problem that is super important. Let's be clear, this is a super important problem for computer vision. Specifically, the self-driving car thing I think is the most important thing that comes to mind. Obviously, it's important for drones and a whole host of other problems.

I guess my point is just that there are problems that are great for crowd-sourcing and there are other problems that are not, and I'm wondering what is the general case of a problem that is a good fit for the crowd-sourced solution?

[0:40:09.3] FP: In my opinion, every problem for which you don't have a solution is good for crowd-sourcing. Whether you know you are looking for a faster sorting algorithm, right? If you look like for — Of course, as you said, machine learning doesn't have an analytic solution, so there's going to be a lot of people trying to make ensemble of different strategies which in general are the one who wins this competition. Different flavors, different nuances until someone gets best results.

Even analytical challenges will be finding — There are prizes in physics that have been there for 100 years. I don't remember the names now, but for example there are these popular. If you prove a theorem then you get a money for it. Every problem which doesn't — This is like — The way I think about it, it's a different business model. Before, it was like, "Okay, we'll try to solve the problem our self within the company."

Now, what happens is that, "Wait, if we don't solve the problem, if our 10 engineers don't solve the problem, we're not going to make money out of it. We're not going to be able to produce to make the product that we want." How about, in first place, we shout out the knowledge that we have to the others to solve the problem. Next, we basically provide the data and all the possible tools in order for the crowd to find a solution for it, to find a better solution.

This is why for example, I think, this is related to all these APIs, like TensoFlow being like publicly available to everybody. Why before a company make their best tools available to everybody for free? Because, now, crowd-sourcing is a different business model, I think. It's a different way of carrying up business.

[0:42:06.8] JM: I was doing a bunch of shows where I was asking people the question of whether we should be worried about AI risk or whether it is something that is like, as Andrew Ng said, the risk of overpopulation on Mars, where you don't need to worry about AI risk, you don't need to worry about the — Or in the near future you shouldn't be worrying about the AI rising up and wreaking havoc on the world.

That question I think it is an interesting question, but what I'm just trying to realize I think is a more pertinent question is I'm looking at these things like the generative convolutional neural networks where you can recreate video based on a large sample. I was looking at some stuff that came out of SIGGRAPH recently or just some videos from projects that are going to be talked to be at SIGGRAPH. I think that's an upcoming conference, but there was something where you can take a large sample of Trump videos and recreate new videos of trump talking. It's quite clear that this is going to become a problem. This is going to be something really confusing in the future where we're going to start to see videos that are really convincing but are manufactured generated videos.

I guess my question is are you starting to have conversations with people about how to maintain the ethics in this type of environment? How to maintain veracity for videos?

[0:43:47.3] FP: I never had this type of conversation actually.

[0:43:49.3] JM: Okay. If you guys don't want to talk about it — I realized this is outside the bounds of your work, but it's just something that's been alarming to me recently.

[0:44:00.3] FP: Yeah, I know you're right. Sometimes I think about it — For example, what's going to be about art? What is like the human touch to art? Art meaning like writing books, making music, generating videos, pleasant videos or something. We're talking about computer vision, but now they're saying neural networks generate music really well. Are we going to need musicians? I don't know how to take it honestly. I'm thinking about it once. I never had talked in the office about it. I'm not personally…

[0:44:37.9] JPT: In my case, I'm on the side of not worried at all, let's say, in the sense that of course our lives will change a lot. Our lives will be much different, but we will adopt and we will survive, of course, to AI, and I wouldn't be worried at all. Of course, as you were saying, maybe I would bring the discussion to a more concrete level and to a more specific level. For instance, we were talking about self-driving cars.

That problem that we were saying whether the machine has to decide whether there's a risk situation, whether to kill directly some pedestrians, three pedestrians and save your life. Or whether to kill yourself and save those three pedestrians. Those are things that are, to me, more tangible, more on the near future, and that we will have to code it somehow and we will have to decide it somehow. That's what can be like really, really more interesting to me to decide.

Whether you take the greater good, let's say, the global good of society versus the private good of your saving your life. What I worry is that it will come down as a lot of things. It will come down to money. I wouldn't like to see it, but it could be depending on the package that you pay or how much you pay, the changes of your surviving will be more or less.

Right now we have it somehow because we have better cars. Theoretically, with better cars, you have less chances of dying, let's say. But now we will have it more, let's say, coded somehow into the neural network and that would at least will create a lot of discussion, and I would say very interesting discussions.

[0:46:28.9] JM: When we're talking about the stuff like self-driving cars and image recognition, does the public research field and the university research field, does that represent the cutting edge or is the cutting edge of this technology sitting behind closed doors in a corporation?

[0:46:47.4] FP: I think the cutting edge is in the resource actually and companies are following up on that. It has to be like this for many years. I guess the gap is closing right now in the sense that the companies are closer to research, but the cutting edge you see at conferences are not within the closed door of companies. I think in the closed door companies, what you see, it's a lot of optimization and making thing works, which is as important as making the initial research I guess, but it's a bit different.

[0:47:21.4] JPT: I would say that right now these line is being blurred a lot by a lot of companies investing in open-research, but I agree with Federico that, let's say, the thing that will work in 10 years times in companies are the ones that people are — In universities, in research in general are researching today.

[0:47:44.0] JM: I guess let's close off. Why do you guys are most excited about in this space of video object segmentation, being able to track images in video overtime. This is a massive tool if we get it right. What are the applications that you're most excited about and what else in the space are you most excited about?

[0:48:08.2] FP: Applications, of course, like low-level — Like video segmentation sort of the purpose of other high-level application, like self-driving cars, like video surveillance for example, like actual recognition. There are many computer visual related task for which video object segmentation might be used as a pre-processing step.

For example, from a visual perspective point — Sorry. For example, for the visual effects industry, video segmentation is a fundamental tool which would cut the cost of rotoscoping.

Rotoscoping is basically the task of artist to basically separate, again, foreground object like actors from the background. So having a reliable video object segmentation algorithm would basically allow the movie industry to get rid, for example, of the virtual green screen.

That will be like a massive improvement for the movie industry. It would able to cut the cost significantly.

[0:49:13.9] JM: What about you other guys?

[0:49:16.1] JPT: For me, of course — Federica comes from Disney Research. Yeah, he is aware of all the industry problems. In my case, I am, as we said, more interested in self-driving cars. As you said, it's a clear situation in which right now it's a hot topic I would say. Also, robotics. Being able to a robot to really grasp so you can tell for instance an assistive robot for handicapped people or let's say, "Go grab that cup of coffee." Really being able, with computer vision, to really know exactly where that object is and segment it and track it overtime. That could be a breakthrough also in robotics, to lower the prices of the sensors that with just a webcam or just a camera you could do all that. Definitely a lot of exciting applications.

[0:50:18.9] JM: All right Guys. I want to thank you for coming on Software Engineering Daily, it's been great talking about The DAVIS Challenge and video object segmentation and I look forward to seeing the results of this year's competition.

[0:50:30.0] FP: Thank you. Thank you very much for having us.

[0:50:32.0] JM: Okay. Great. All right. Thanks guys. This is great. I think there was — Wait. There was one of you that didn't answer the last question, but I don't — Is that okay? Did you want to answer it?

[0:50:44.4] SG: No. I kind to follow up a little bit on Jordi's answer. What [inaudible 0:50:49.8] for grasping objects is crucial to have a pixel wise annotation and non judgement inbox. It would be difficult. Also, for the drones, that in order to track people, to follow people in the videos. Also, they have to do any kind of manipulation with also a robotic arm for repairs and [inaudible 0:51:13.1] or something like that.

[0:51:15.8] JM: Right. That's going to be huge. I can't wait till a robot can just fix my sink. Anyway, cool.

[END OF INTERVIEW]

[0:51:31.8] JM: You have a full time engineering job. You work on back-end systems of front-end web development, but the device that you interact with the most is your smartphone and you want to know how to program it. You could wade through online resources and create your own curriculum from the tutorials and the code snippets that you find online, but there is a more efficient option than teaching yourself.

If you want to learn mobile development from great instructors for free, check out CodePath. CodePath is an 8-week iOS and android development class for professional engineers who are looking to build a new skill. CodePath has free evening classes for dedicated experienced engineers and designers. I could personally vouch for the effectiveness of the CodePath program because I just hired someone full-time from CodePath to work on my company Adforprize. He was a talented engineer before he joined CodePath, but the free classes that CodePath offered him allowed him to develop a new skill, which was mobile development.

With that in mind, if you're looking for talented mobile developers for your company, CodePath is also something you should check out. Whether you're an engineer who's looking to retrain as a mobile developer or if you're looking to hire mobile engineers, go to codepath.com to learn more. You could also listen to my interview with Nathan Esquenazi of CodePath to learn more, and thanks to the team at CodePath for sponsoring Software Engineering Daily and for providing a platform that is useful to the software community.

[END]