ICCV DAILY

Friday 27

International Conference on Computer Vision
Venice, Italy - 22-29 October 2017

Presenting Work by:
Emanuela Haller
Marius Leordeanu

Expo:
Second Spectrum

Interview with the General Chair:
Gérard Medioni

Today's Picks by:
Christian Unger

Women in Computer Vision:
Vicky Kalogeiton

In cooperation with
Computer Vision News
The Magazine of The Algorithm community

A publication by
RSIP VISION

For today, Friday 27

Dr. Christian Unger finished his PhD on stereo vision in 2013. He works at the manufacturer of the ultimate driving machine (the BMW Group) in the autonomous driving platform where he coordinates the computer vision topics.

"I am generally highly interested in stereo vision, 3D reconstruction, optical flow and methods for object recognition. To this end, I would like to encourage everyone to work on these still challenging topics, particularly in the context of vehicles moving through adverse environmental conditions. "

Christian's picks of the day:

- **Morning**

O7-5: DCTM: Discrete-Continuous Transformation Matching for Semantic Flow
P7-20: 2D-Driven 3D Object Detection in RGB-D Images

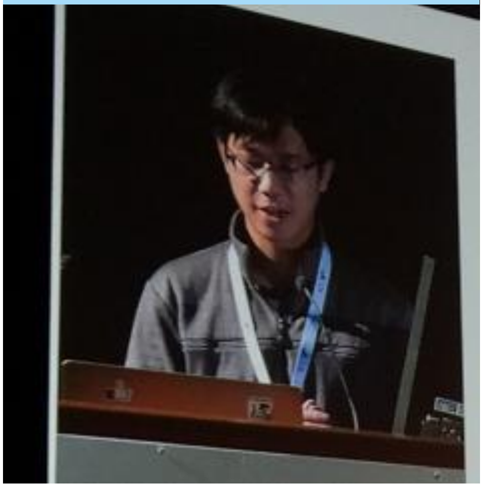
- **Afternoon**

S7-1: Quantitative Evaluation of Confidence Measures in a Machine Learning World
P8-22: Camera Calibration by Global Constraints on the Motion of Silhouettes
P8-40: Long Short-Term Memory Kalman Filters: Recurrent Neural Estimators for Pose Regularization

"The ICCV is a great conference to keep in touch with current international research trends and former colleagues. I really enjoy all the novel concepts and fresh, inspiring ideas. I also would like to thank the organizers for the wonderful weather (which somewhat compensates for the "delightful" and "precious" lunch boxes ☺). Venice is an impressive place!"

Christian's Picks

Kaiming He - FAIR Best Paper Award MASK R-CNN



Gérard Medioni

Marius Leordeanu Emanuela Haller

AutoDIAL

Expo: Second Spectrum

Julya Peyre

Women in Computer Vision Vicky Kalogeiton

ICCV Daily
Publisher: **RSIP Vision**
Copyright: **RSIP Vision**

Our editorial choices are fully independent from ICCV and its organizers,

Gérard Médioni is a professor at the University of Southern California, Director of Research at Amazon, as well as General Chair at ICCV 2017.

We are in the middle of a week rich of technology and talent. What are your initial thoughts about the first days of the conference?

It is amazing. It's an amazing conference. To tell you the truth, we were planning to have 1,800 people, maximum. We were in Chile last year, and we had 1,500 people. We said, because we are going to be in Europe, we're going to have 1,500 people. We have over 3,100 people today. It is an amazing conference. The other part is that we have a large number of companies that are coming to attend. If you go back about four or five years for ICCV, ECCV, and CVPR, we had some companies, but never to that scale. What it shows is that our field has matured to the point where we can create real solutions with computer vision. That is why these companies are coming and looking for talent, showcasing their abilities to deliver real solutions.

Were these two successes, the large attendance and the large response from companies joining the expo, something that you expected or were you surprised?

It was a surprise. We were not planning to have that many. We made sure that we could expand, that we would have the room to expand, but even then we were surprised. We could not accommodate 4,000 people, for example. We were hoping no more than 3,200 people would show up.

Otherwise, we would have to cap the attendance which we don't want to do. It's very painful for an organizer to say, *"Sorry, you cannot come because we don't have any more space."*

Actually, the hotel will tell them to sleep in the street.

Absolutely! [*both laugh*]

So we were surprised. We were surprised both by the number of attendees and also by the number of companies that came and wanted to sponsor an exhibit. We had one company, I'm not going to name it, that came about a week before it started and said, *"Can I please still join?"* Of course, we did accommodate.

You have been at this kind of conference for many years. What strikes you about the people you are

meeting now which is different from before?

So there is a change in the number of people because now we have many more attendees that are not presenting. This used to be a conference where everybody was presenting something. What has changed now is that we have a number of companies that are coming that have a booth, and they are presenting. Their role here is not to come and present research, but rather to attract talent. That has changed. What I've seen also is a difference in the way we present results. Today we have large datasets where you can measure progress. I'm kind of a pioneer from back in the old days of computer vision when you created your own datasets. You presented your results, and people had to evaluate the quality and the value of it in a very non-scientific way, in a sense. Today, what you have are datasets. You can measure 50% this or 10% that. That is a measurable difference which allows us to say, *"Algorithm A versus algorithm B: I know how these two can compare, and this is what we should be going with."* That has allowed the field to move much faster than it moved before.

Of all the things that you've heard in the last few days, is there one "wow" moment that struck you in particular?

Ah, that's a difficult question. I like, very much, the oral papers. I'm not going to cite them. I think the award papers are very much deserving. Every one of the award papers was different and brought a new perspective, a new idea that was really not known before. That's why they got the award. I think the award committee did an excellent job of selecting the papers. That said, it is sometimes difficult to evaluate the impact of a paper that is being published. There are awards now for papers that have stood the test of time. That, to me, is valuable, to look back, not just at this conference, but let's go back ten years. Ten years ago, which paper actually made a difference that people are citing today and are using today?

What impresses you when you see a presentation or you read a paper?

There are different ways that I can be surprised. One is a result that I didn't think we could achieve. There was a paper from MIT that talked about looking around the corner.

Of course, by Katie Bouman!

That's right. That's a surprise. That's something where you say, *"I didn't think that you could do something with these types of images."* Again, they show that, yes, there is information that we can extract. That's one type of surprise. The other type of surprise is producing a rich caption from an image or from a video. Five or seven years ago, we would have said that there's no way we can do something like that. The field has advanced tremendously in that respect.

When did you start organizing ICCV?

It was four years ago. I have an interesting story to tell you. There were two competing proposals. One was for Paris, and one was for Venice. The first presentation was for Paris. This was a very good presentation. Then I came up, a Frenchman, and said, "You must be wondering why your Frenchman is here trying to convince you to go to Venice instead of Paris." I said, "Paris is a wonderful city. I think the proposal for Paris is excellent. I think we should go to Paris at some point. However, I'm here to tell you we should go to Venice because Venice is sinking. If we don't go to Venice then we might never be able to go." [*both laugh*]

It's a sad and funny story at the same time.

Absolutely. I don't know whether this had an influence or not, but basically that was my selling point. From a scientific point of view, those proposals were outstanding. There was very little difference between them, but I felt that coming to Venice was a unique opportunity that was time bound, and therefore, that's possibly what made the difference. There wasn't much of a difference between the votes, probably 10-20 votes between the two proposals. Venice won.

Four years later, can you tell us what challenges you faced that you did not expect in organizing a conference like this?

I normally organize CVPR conferences which are in the US. With a non-US company, doing it remotely, we have to rely very heavily on the local organizer such as Rita and Marcello, who did an outstanding job. But there is a cultural difference. For example, we signed the contract with the venue only a couple of months ago. This is something that wouldn't be possible in the US. In the US, either you have the contract signed or you don't have it. In Italy, it is understood that we have it. The fact that we didn't sign it isn't as important. There is this cultural difference.

It was nearly a miracle this year that CVPR could move from Puerto Rico to Hawaii with a very short notice.

Right, and that was okay. It was very nicely done. The better example is the one in Las Vegas. It was originally scheduled for Seattle. The reason we didn't go to Seattle is because we didn't sign on time, and we lost the venue. That's why here, I was very worried that we didn't sign! My Italian colleagues kept saying, *"Don't worry! Don't worry! It's okay."* [*both laugh*]

This is how we do it!

They said, *"No it is not signed, but it* *will be signed. Don't worry about it!"*

Va bene, va bene!

Exactly! [*both laugh*]

> *"The venue itself is fabulous!"*

That was the challenging part. What was the positive takeaway of organizing this conference?

First of all, working with the professional conference organizer went very smoothly. They were very, very professional. Cristiana and the entire organization were absolutely outstanding. It is excellent.

All of them, Federica in particular.

You can see it in the way they organize, the way the badges are checked and printed, the way they organize the lunches. It is very smooth. The best thing with an organization is when you don't notice it. This is exactly what happened. The venue itself is fabulous! This is where the film festival occurs. The room is splendid. We have this audio/visual link for the overflow room.

And the sea in front of us!

And the weather that we ordered!

[*both laugh*]

That is absolutely wonderful. In fact, we are sitting outside in front of the venue.

And very much enjoying a chat by the sea. What advice do you have to any young student attending their first conference here at ICCV?

My main advice is don't be overwhelmed. It is very difficult to come here. There are so many talks, so many workshops. We have 44 workshops here. It is very easy to say, "*Oh my God, I will never be able to make an impact in this field.*" That is not true. Make sure you choose a topic. You understand it. You read what has been. Make sure that you can actually advance the state of the art. Instead of being in the audience, you are going to be presenting to the audience on stage.

> *"You are lucky to be in a field that has exploded in the past five years..."*

This is my advice. And know that you are lucky to be in a field that has exploded in the past five years. This is a field that five years, seven years ago, if you said I work in computer vision, they would say, "*Oh, okay, nice.*" Now, people say, "*Oh! You know what, now there is a plethora of job offers.*" People are looking for talent in this area. There are so many opportunities from self-driving cars to internet processes. Everybody is consuming images, but now the difference is, instead of consuming pixels, we want to get information out of the images. This is what our field is doing, and I think it's going to continue having a great influence in the world.

Weakly-Supervised Learning of Visual Relations

Julia Peyre is a PhD student at Inria Paris, supervised by Josef Sivic, Ivan Laptev and Cordelia Schmid. She speaks to us about her upcoming oral and poster.

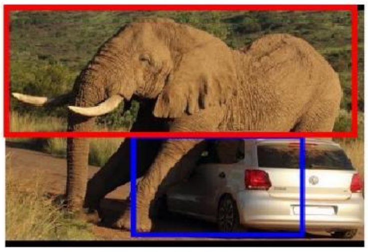
To begin, we ask Julia about working with **Cordelia**: *"It is very nice. She is a very efficient person. Nothing is random: she goes straight to the point. We are never losing time."*

This work is called **Weakly-Supervised Learning of Visual Relations** and its goal is to learn relations between objects in images, using only weak supervision for the relation. Typically, the input of this method at training time will be an image with image-level triplets of the form: subject, predicate, object. For example, there is a person riding a horse in an image, but they do not know the localisation of the objects. They will train the method to learn a classifier for the predicate – riding, in this example – using only this kind of supervision.

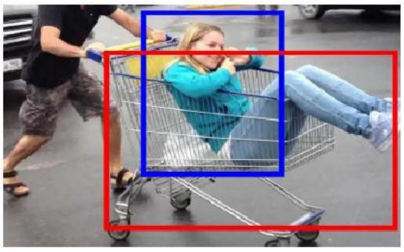
This task was first introduced at ECCV 2016 in a paper called Visual Relationship Detection with Language Priors, by **Cewu Lu**, **Ranjay Krishna**, **Michael Bernstein** and **Fei-Fei Li**. That paper solved the task described above, to detect the object in a certain relation in images; however, at the time of publication, it was not addressed with weak supervision. That is the novelty of this work.

The development came about because the team had been interested in relations between objects for Julia's thesis. In their lab at **Inria**, they had been working with weak supervision, so it was natural to think about doing this task with weak supervision. Julia adds that it is important to do this with weak supervision because it is a very challenging problem to get annotations at box-level for the relations.

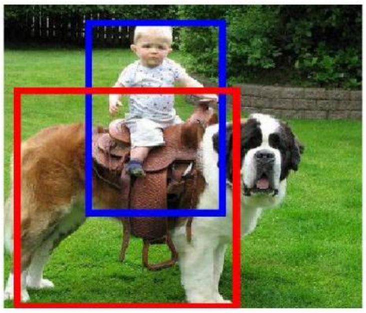
Julia explains: *"If you take natural images you will have a lot of objects in these images and the objects will have many different interactions. If you want*



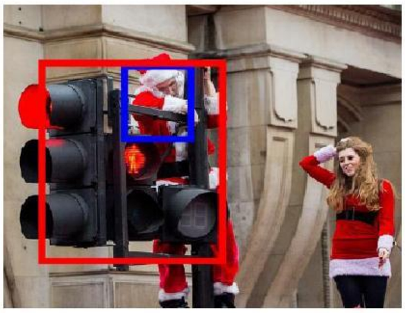
car under elephant



person in cart



person ride dog



person on top of traffic light

to learn with full supervision these kinds of relations, you would have to annotate all the relations between the objects in an image. Getting this kind of annotation is very expensive, because the total number of annotations you would have to get for one image is n2, where n is the number of objects in your image."

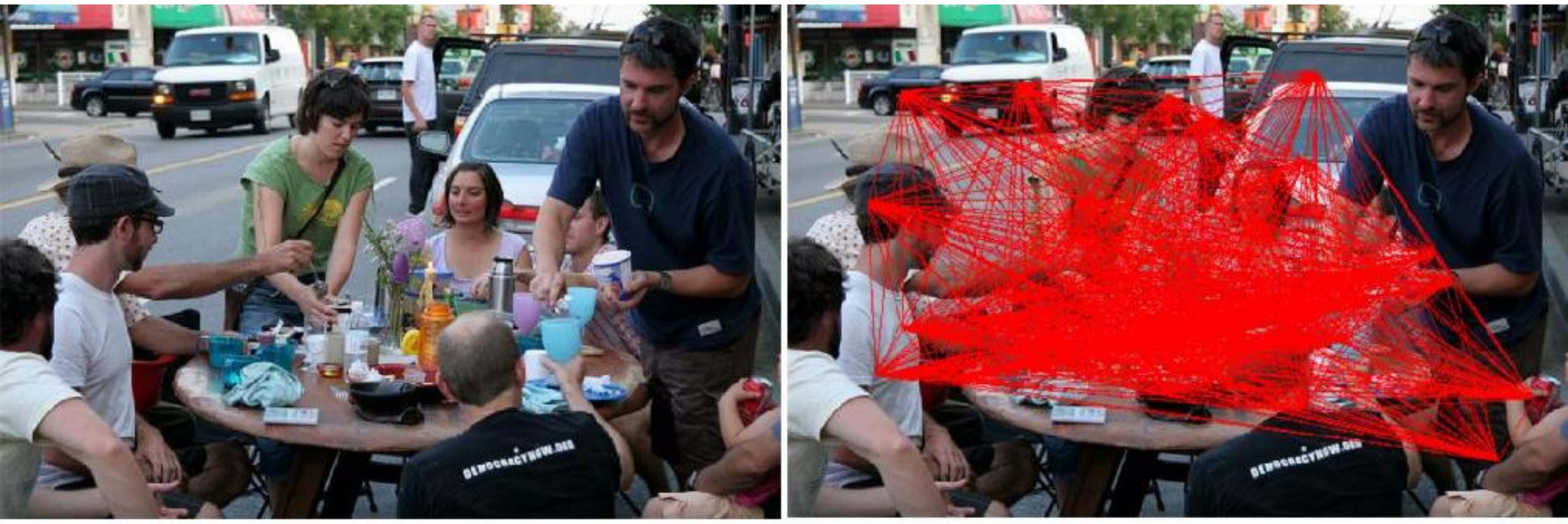
The main benefit of this is that they just require annotation at image level, so people don't have to draw the boxes between all objects and annotate the relations for all of those pairs of objects.

Julia tells us their method is in two stages. The first stage is to get candidate objects for images. For this, they use a standard object detector. Then they have these candidate objects as proposals and want to learn the relations between them. For this, they use a method called discriminative clustering, which is a very simple framework developed by **Francis Bach** and **Zaïd Harchaoui**. It is a very flexible method which allows them to incorporate constraints very simply.

Julia says the next step is to move towards using more natural language. Right now, they require image-level triplets. They are constraining annotation to be in the form of triplets inside a limited vocabulary – a fixed vocabulary for object and a fixed vocabulary for relation – so the next step is to learn directly from captions. On the internet, if you want to use web data, you will encounter natural language, not triplets.

Julia concludes by saying: "*I would like to advertise a new dataset that we introduced which is called UnRel for unusual relations. This dataset also answers the difficulty to get annotations at box-level, but this time at test time, because you encounter a lot of missing annotation for evaluation, that would introduce noise at evaluation. To solve this problem of missing annotations at test time, we introduce a dataset of unusual relations. For example, a dog riding a bike or a car in trees. The advantage of using this dataset for evaluation is that you will have a reduced level of noise in evaluation. You can now evaluate with retrieval, without worrying about the problem of missing annotation.*"

If you want to learn more about Julia's work, come today (Friday) to her oral at 13:30 and her poster at 15:00.



Difficulty of getting box-level annotations for relations

Second Spectrum is exhibiting at the ICCV2017 EXPO. It is a sports analytics company with around 160 employees, based in Los Angeles. It also has a small team in Lausanne, Switzerland and an office in Shanghai, China.

Horesh Ben Shitrit, manager of the Computer Vision team at **Second Spectrum**, tells us more about the company:

"We are doing sports analytics from computer vision, machine learning and artificial intelligence. We use multiple cameras that we install in the courts, then we analyse the position of the players and the ball. We get the position of everything that is moving in the court in almost real time and then do all the analytics on top of that. We are analysing all the events that are happening. All the actions on the defense and offense. For example, in basketball, we have the pick and roll plays, blocks, assists, shots. All the probabilities and efficiencies of all the players."

Horesh says that every sport has its own characteristics. In basketball, it can be challenging because there are a lot of occlusions. In football, it can be challenging because it is an outdoor scenario, with weather conditions like rain, fog, snow, and even smoke from the audience to deal with. However, for most games it goes very well, and they can track all the players, the ball and the referees.

Six years ago, at ICCV 2011, Horesh published a paper on multi-object tracking that showed the basketball player tracking they had developed in the academic setting. He says it worked nicely, but it was just a lab prototype. Six years later, Second Spectrum has a working system that is the most prestigious in the field and is the official tracking provider of the NBA. It has been an amazing journey for them.

"We get the position of everything that is moving in the court in almost real time and then do all the analytics on top of that"

Horesh tells us they also have a product for coaches which allows them to analyse their games. They can see themselves and their opponents, including all the video of different players and actions.

ICCV Daily was told also about what to expect from Second Spectrum in the near future: “Right now, everybody turns on the television and sees exactly the same game through the eyes of the broadcasters. We are going to change it so that every person sitting at home sees it in a different way. If you are an amateur and would like to understand the game itself, you will get an explanation about how the formation is working. If you are an expert, you would see it through the eyes of a coach, and get more statistics and understanding about tactics. If you are watching it with kids, you are going to see it in a Snapchat-style with some augmentation offering cool stuff. This is the future. We are going to personalise the user experience.”

Horesh concludes by telling us that Second Spectrum is looking for interns. They have summer internship positions for Masters students or Ph.D. students. After that, they can join the team and continue to do computer vision business for sports analytics.

“Second Spectrum is looking for interns. They have summer internship positions for Masters students or Ph.D. students. After that, they can join the team and continue to do computer vision business for sports analytics”

“This is the future. We are going to personalise the user experience.”

Friday
ICCV DAILY

AutoDIAL: Automatic DomaIn Alignment Layers

Elisa Ricci is Assistant Professor at the Department of Engineering at the University of Perugia and a researcher in the Technologies of Vision group at Fondazione Bruno Kessler.

Elisa will present a poster co-authored with Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, and Samuel Rota Bulò. It is joint work between Fondazione Bruno Kessler, Mapillary Research and La Sapienza University of Rome.

The work tackles domain adaptation, which is an interesting and challenging problem in computer vision. There are already many works on the subject of domain adaptation and recently on deep domain adaptation – that is how to handle domain shift in deep networks. Domain shift means that there are huge discrepancies between your training data and your test setting.

The main advantage of the team's approach is simplicity. They have layers called domain alignment layers that can be embedded in any architecture to cope with this domain shift problem and reduce it. They have a method that can be implemented in a few lines of code which outperforms the state of the art in one of the most challenging problems in the community.

From left:
Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci and Samuel Rota Bulò

“The main problem in domain adaptation is how to measure domain discrepancy and how to cope with it”

Elisa explains that the main problem in domain adaptation is how to measure domain discrepancy and how to cope with it. There are many strategies for deep domain adaptation, but the question is: which one is the best? Their strategy of adaptive batch normalisation has proved to be very effective, but they know that there are others.

The idea is to adapt the batch normalisation layers, that are very popular layers in convolutional neural networks, and use different statistics for the source and target domain. Elisa says: “I think one interesting thing about our approach is that it is the first method where at each layer of the network, the network automatically chooses how much to adapt. This, to my knowledge, has not been done in previous work and I think it is one of the strengths of our approach. This leads to very good experimental results.”

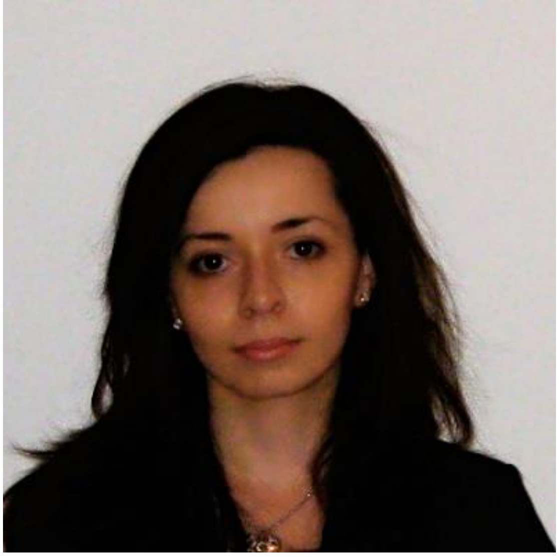
One feature the team would like to add to the method is a way to better measure the distance between the distributions. They are using a simple way to model distributions, considering just the mean and variance, but Elisa says they could explore using high-order statistics, histogram matching, or another more sophisticated way to change these layers. They want to keep the idea of automatically choosing at each level of the network how much to adapt.

As for next steps, Elisa tells us that there is more to investigate: “*This work can be extended to a wide range of applications. Now, we use a common benchmark based on object recognition, but I think the work can be extended to many other problems. There are a lot of problems where you need to cope with domain shift. For instance, you have different illumination conditions, you have people moving from different cameras. Investigating the applicative point of view, it is important. We should move there.*”

If you want to learn more about this work, visit Elisa’s poster today (Friday) 10:30 at Sala Mosaici 2.

“The work can be extended to many other problems. There are a lot of problems where you need to cope with domain shift”

Unsupervised Object Segmentation in Video by Efficient Selection of Highly Probable Positive Features



Marius Leordeanu is Associate Professor at the University Politehnica of Bucharest as well as a senior researcher at the Institute of Mathematics of the Romanian Academy. He supervises Emanuela Haller, who is currently a PhD student at the University Politehnica of Bucharest. Marius and Emanuela will present their poster today (Friday) at ICCV2017.

Their work focuses on video understanding. Together, they published a paper on discovering objects in videos in an unsupervised fashion. They have formulated a solution for efficient and fast object discovery with a much different approach than many of the other current papers. Their paper proposes a general approach that could work with any kind of classifier and will benefit from current neural nets techniques. It explores object discovery using a simple solution based on common sense and the basic difference between a foreground and background object.

They took a completely different approach to unsupervised learning which no one else has done before. They came up with a solution for learning from highly probable positive features. This differs from current approaches that try to find the differences between the object and the background. Instead, they took a new approach that tries to learn what the object looks like. By knowing the object well enough, they can differentiate the object from the rest of the background.

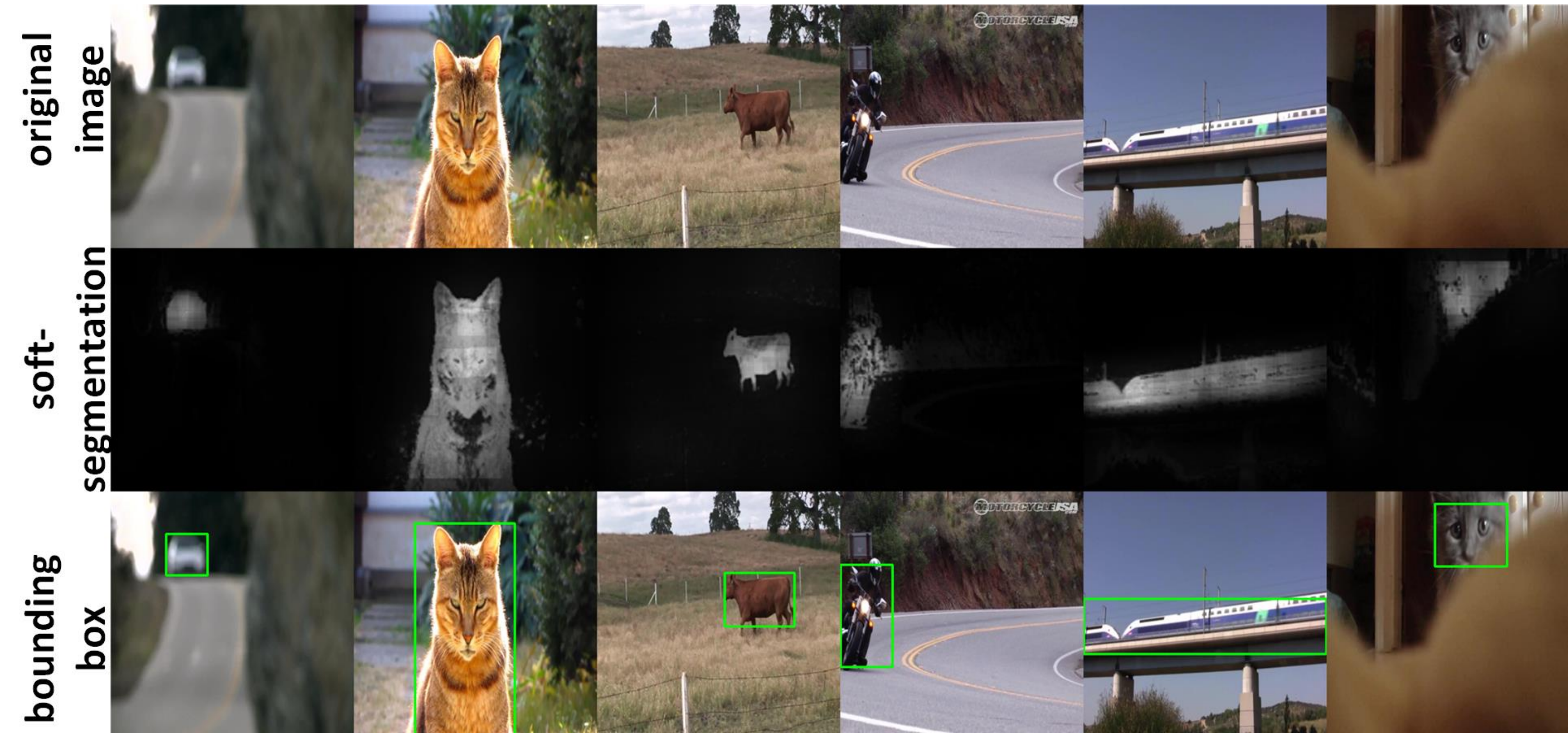
When asked about the challenges of the work, Marius revealed, “I think that this work is showing a very interesting secret that we have discovered regarding unsupervised learning.” In their view, unsupervised learning can be done in a video because it has spatial and temporal consistencies.

He then added, *"There is something that nature tries to tell us in video. If we just look at random pictures from the internet, it's very hard to discover things in an unsupervised way. Whereas in video, we have this consistency over space and time. We have co-currencies of certain patterns. We can take advantage of them in order to discover what might be related to a certain object."*

They took a very general approach that could work well with deep learning as well as with any type of classifier. In supervised learning, most works go towards the hard cases of positive and negative examples. They came up with a completely different idea and decided to start with the easy cases where they could select with high precision, positive samples.

Marius explained, *"We can pretend that they are positive samples, even though the recall is low, which means we don't get all of them, but we get some of them using certain cues. Then we learn a classifier, and we also have a theoretical result which proves that this will be almost equivalent to having the full set of positives. This classifier, the next iteration, will increase the recall while keeping the precision high. This means that now, we have the positives, and we can come up with a better classifier. In this way, the next iteration classifier is richer, stronger, and so on."*

"There is something that nature tries to tell us in video. If we just look at random pictures from the internet, it's very hard to discover things in an unsupervised way. Whereas in video, we have this consistency over space and time. "



In the end, they hope to increase the recall so the precision recall curve will increase. This allows them to approach more difficult cases. They could bring in very powerful systems and classifiers based on deep learning and convolutional networks.

On the practical side, their method has the advantage of working very quickly. Even in math labs, it is about three or four frames per second, and this could easily be done in real time in C++ implementation. It also does not use any pre-learned features while still obtaining state of the art results when compared to other approaches which may look more complex.

“Their method gives them high precision and good quality features that they can harvest”

In essence, they try to focus on the problem, not on a specific system or specific technique. They want to find out what they can learn by simply watching videos. The next step would be to achieve better performance and more complex datasets. Because the datasets are constantly changing in this field, they want to employ all sorts of neural nets and test it on the most complex datasets available. That means bringing the state of the art to the next level. Marius feels extremely optimistic about this idea. Together with Emanuela, he plans to dedicate at least three to five years to this idea.

When asked why she loves this work so much, Emanuela responds, “I think a solution to this problem would describe how we are understanding the world. It’s how children learn their world basically. They see the object move in front of them, and they learn how to recognize it the next time.”

She explains how this method teaches ways to find an object in the next frame without knowing what the object is. Whether a car or something else, it doesn’t matter. You just need to know that it is an object, and you learn how it acts in real life.

Marius compares their approach to the analogy of a fisherman. A fisherman must use certain cues on how to catch a fish without knowing everything about the fish or where to find it. Imagine a fisherman in a river for the first time looking into the water observing for movements. Perhaps he throws something into the river to help find the fish. After he catches the fish, he looks at the fish and learns even more about it.

“They want to find out what they can learn by simply watching videos”

Through this learning process, the fisherman gains a better understanding of the fish, its behaviors, and cues to look for to find it. In the end, it becomes an easy task knowing much more about how to catch the fish.

In a similar way, their method gives them high precision and good quality features that they can harvest. With every iteration, it improves. At the end, they obtain really hard positive and negative cases and reach the human level performance and beyond.

To find out more about Marius and Emanuela’s work, visit their poster today (Friday) at ICCV 2017.

Women in Computer Vision

Vicky, how is it working with Vittorio and Cordelia?

They are both extremely admirable people. More importantly, they are even more admirable as researchers, as you already know. Vitto is an extremely charismatic person that is fun and inspiring. He has a passion for everything. When he is excited, he has this ability, this charisma, to actually transmit his excitement to everybody else. That is excellent for young, junior researchers like myself who are struggling, struggling, and struggling when nothing works, you have little strength to cheer yourself up. This helps a lot and makes the miserable days, weeks, or months when nothing works much happier. At least you have a perspective. You have a positive attitude. Cordelia, on the other hand, is excellent in understanding, in visualizing, in thinking through things, which is an amazing ability that I would love to inherit at some point. When you tell her something simple,

Vicky Kalogeiton recently completed her PhD under the supervision of Cordelia Schmid and Vittorio Ferrari at the University of Edinburgh and INRIA.

she can actually take it to the next level. She can transform it into a completely different problem that, according to her, is going to be close to what you said in the beginning, but it's not even the same thing sometimes. [laughs] It's going to be so far away in a sense that you would not have been able to predict it.

What would you like to inherit from Vittorio then?

Obviously, his passion and his intelligence. It's an excellent combination. Vitto's mind works faster than hundreds of people.

Do you really feel like you don't have the same passion as Vitto? I saw you at your poster today. You seem very passionate about your work. Am I wrong?

I am passionate about my work, but I guess every PhD student is.

Well, you're not a PhD student anymore! [*both laugh*]

Nice trap!

Can you tell us about the poster that you presented today?

It is an action-tubelet detector for spatio-temporal action localization. In this work, we deal with the spatio-temporal action localization problem. That is exactly what its name suggests.

Localizing when and where the actions take place in a video. For example, you have somebody that is diving in this amazing swimming pool and you want to find out where exactly this human is, spatially on the frame-level, and when his action starts and ends temporally. Until today, state of the art works focused on tackling the problem more at a frame-level. They use per frame detectors that detect the actions at a frame level, then they link the actions over time to create spatio-temporal tubes. There is a lot of good in this method and there have been improvements over the years, but it has some basic disadvantages. It doesn't exploit the temporal continuity of videos. Imagine somebody in this position. Am I sitting down or standing up?

I think you are standing up with your knees bent.

I am sitting down right now.

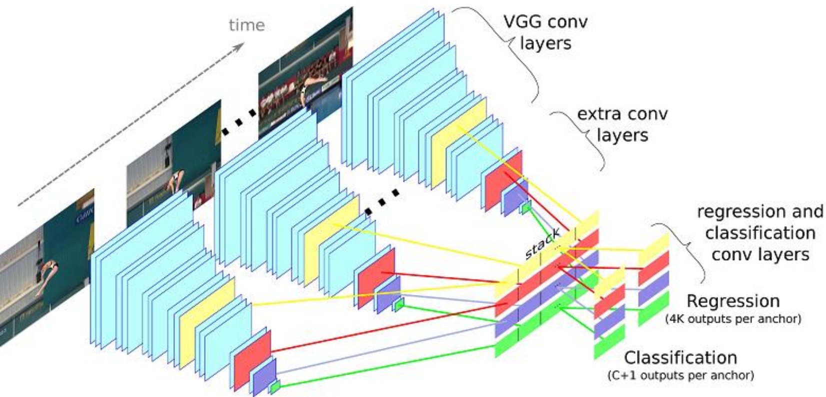
You are. What's the story?

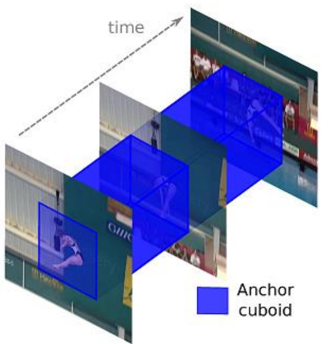
We propose to surpass this limitation and instead of working on a frame level, we work on a sequence of frames. We propose an action-tubelet detector that takes a simple sequence of frames and outputs tubelets. In the way that standard object detectors work, we extend the anchor boxes of standard detectors to anchor cuboids that have a fixed spatial extent over time. We regress the anchor cuboids in order to follow the movement of the actor. We try to say, where the actor is, it will be like this, going up and down. This is the regression part, and the classification – which basically means to put a label on what somebody is doing – like running, kicking a ball, or any action label. To do so, given sequences of K frames, we deploy K parallel streams, one for each of the input frames. We learn our action tubelet detector jointly for all parallel streams, where the weights are shared amongst all the streams. In the end, we concatenate the features coming from each stream, then we classify and regress the anchor cuboid to produce the tubelet, which is what we want.

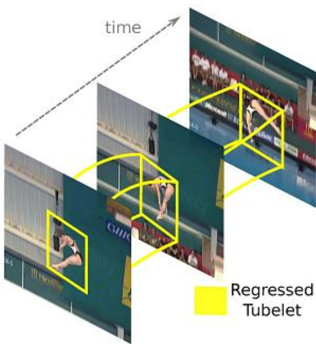
What is interesting there is that the features are learned jointly; however, the regression is done per frame from features that are learned jointly over the whole sequence. The classification, putting the action label into the sequence, is consistent over the whole tubelet. That enforces consistency of the actions over time.

What feature would you add to the model that it doesn't have today?

I could tell you millions of them! What matters though is what I think people who work on action localization or







action recognition would like to do in the long run, which is think in longer-term relationships between video frames. Video doesn't consist of one frame, two frames – that is a very constrained environment – so I assume that the long-term goal is to process the whole video. Of course, in some tasks, that is already the case. Not yet for action localization though. I assume that this will be the long-term plan. Any component that anybody would add would…or should, or I would like it to [she laughs] lead towards that direction. It can be LSTMs, or end-to-end networks or many other things, but I assume that what matters is the general direction.

You are obviously very passionate about this work. Is that because you are a passionate person or because there is something special about it?

Both! [she laughs]

What is so special for you about this work?

My supervisor Cordelia is a person that thinks ahead, and she transmits this to people around her. This project is a way of moving forward. It is one of the first works that perform localization considering a spatio-

served the same purpose. They used motion, which is concatenation of consecutive frames. Again, taking into account the motion is as an extremely good idea. I assume that what I really like is that part. That it is a step forward. Baby steps, little by little.

Where were you born?

I was born in Athens in Greece, then I studied electrical computer engineering in Greece.

Did you feel like science was your passion?

If I say yes then it's going to be too much, but if I say no then it's going to be why not.

So tell us the truth! [*both laugh*]

It always works better. [*both laugh*]

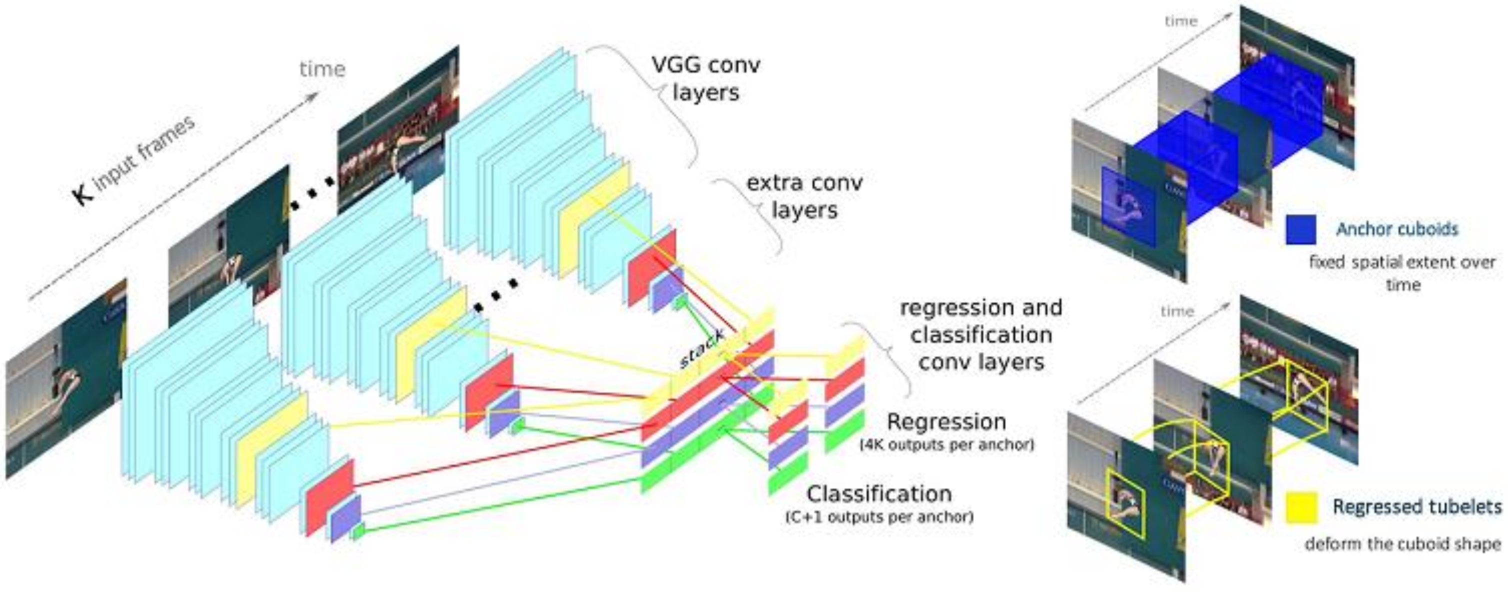
Okay, yes.. Obviously I will tell you the truth. I just need to think a bit more.

Which truth will you say?

[*both laugh again*]

You know, our readers are very smart. They will know! When did you decide to become a researcher in science?

It happened in university. I don't know if it was a decision, a direction, or a desire. It was maybe a combination of



the three, but it happened very early in university, like the first year I think. I was looking around at the people who were interesting or interested in things. They were mostly people who, at that point, were PhD students. I didn't know these PhD students because I was an undergrad in my first year. They were teaching classes. They were the ones that were passionate.

So you thought you would like to be like that?

I guess so, I guess so. I don't know.

Until when were you in Greece?

Until 2013. I did there an undergrad and then a Master's.

Then you said, "*Let's go and find luck somewhere else.*" How did that happen?

Cordelia and Vitto had an open position. I applied to this position for a PhD. It was half at the University of Edinburgh and half at INRIA. That's when Vitto was still at the University of Edinburgh. I applied, and then I spent the first two years in Edinburgh, then the last two years in Grenoble.

Which did you prefer?

Grenoble

Why?

Weather! [*both laugh*]

And food probably…

[Vicky nods, both laugh]

Which one would YOU prefer?

Depends, if I wanted to spend a weekend, I would probably choose Edinburgh.

And if you want to spend a year?

Umm… Paris? [*both laugh*]

Diplomatic. [*both laugh again*]

I don't know Grenoble, but I know Paris....

I love Grenoble.

What is nice in Grenoble?

Weather!

What do you miss the most about Greece?

I guess the spirit of going to sit for hours for coffee, and then after that go for another coffee for hours. [laughs] I don't know if this is only associated with Greece, it's probably mostly the student life, not so much the Greeks. Maybe it was the Greek weather. I loved that. I really think it happens in other countries as well.

Will you go back to Greece one day?

No? Yes? I don't know?

What would you like to do in the future?

I have no idea!

You never thought about it?

No idea!

Well, that's a legitimate answer...

I don't know...

Whatever comes is good?

No, but it's a big decision. It's not whatever may come, but at any given time, I might want something different. Right now in 2017, I want this. Maybe in a few years, I will want something else.

You said before that you could have a bad day, a bad week, or a bad month. How do you prepare for those times so that it doesn't bring you down? Do you have any tips?

I guess you could print a picture of the conference where you would like to submit your work. Put it on your wall or any wall where you'll go so that it is a motivation.

I like that!

Well, everybody has his own tracks.

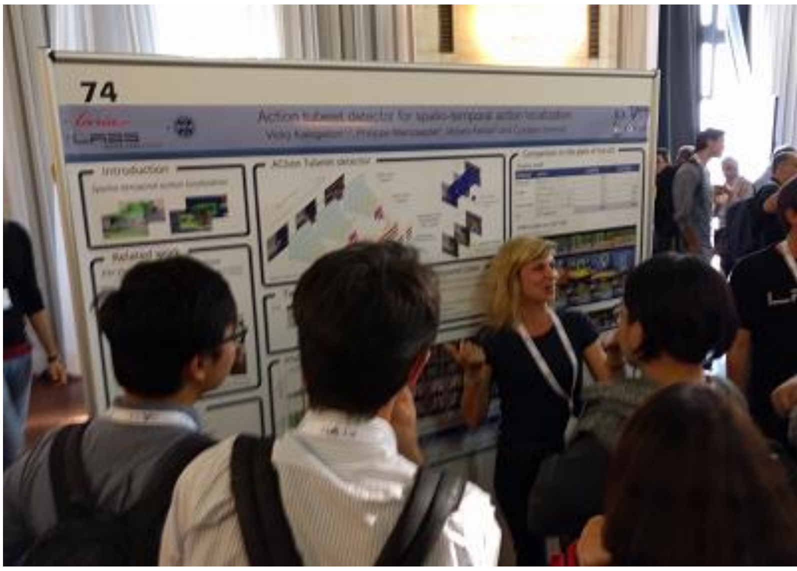
This is why I ask you to share them. When I saw you presenting your poster like you did today, I said to myself "*I'll have what she's having!*" [*both laugh*]

You can spend time with good friends, people who may not necessarily do the same thing as you. You can detach a bit and then go back. I guess these are the standard things, things that hopefully everybody has so that they can relax, maybe walking around or seeing a movie.

Did you ever teach?

No, I haven't.

Would you like to?


74
Action tubelet detector for spatio-temporal action localization

Yes, I would.

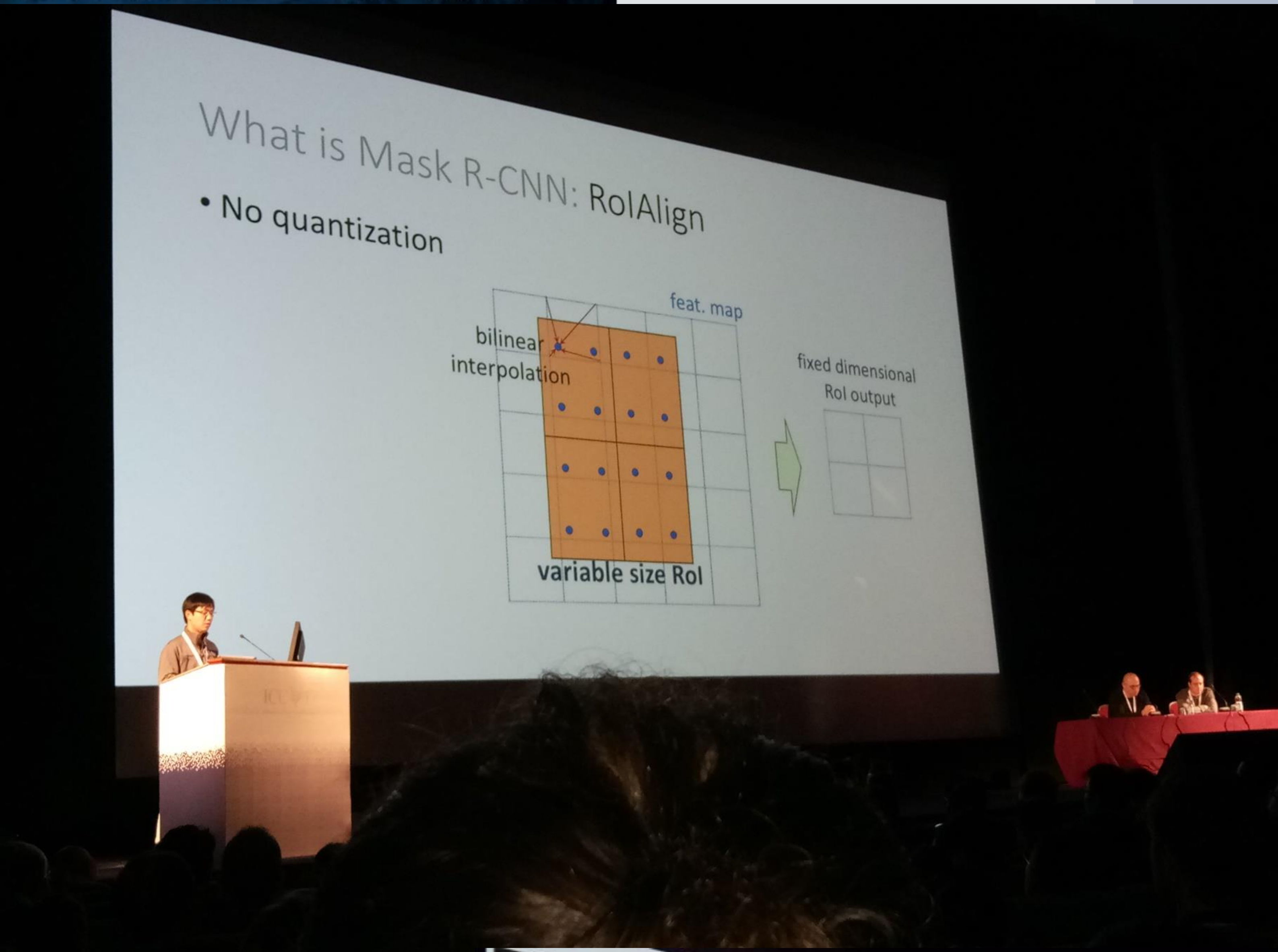
What interests you in teaching?

Well, I assume that this is the standard thing that everybody says, that being a teacher is extremely fulfilling, that it makes you feel good about yourself, that you transmit the knowledge you have. You have the chance to influence others.

Because you had teachers like this?

Yes.

You will be the same!



What is Mask R-CNN: RoIAlign

- No quantization

Kaiming He yesterday at ICCV2017, presenting this year's Best Paper Award winner: Mask R-CNN. Some of us recognized immediately the quality of this work, co-authored with Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Computer Vision News is proud of having published one of the first enthusiastic reviews of this paper, whom we described as "another outstanding work by Kaiming He et al."
Kudos to the FAIR (Facebook AI Research) team!!!

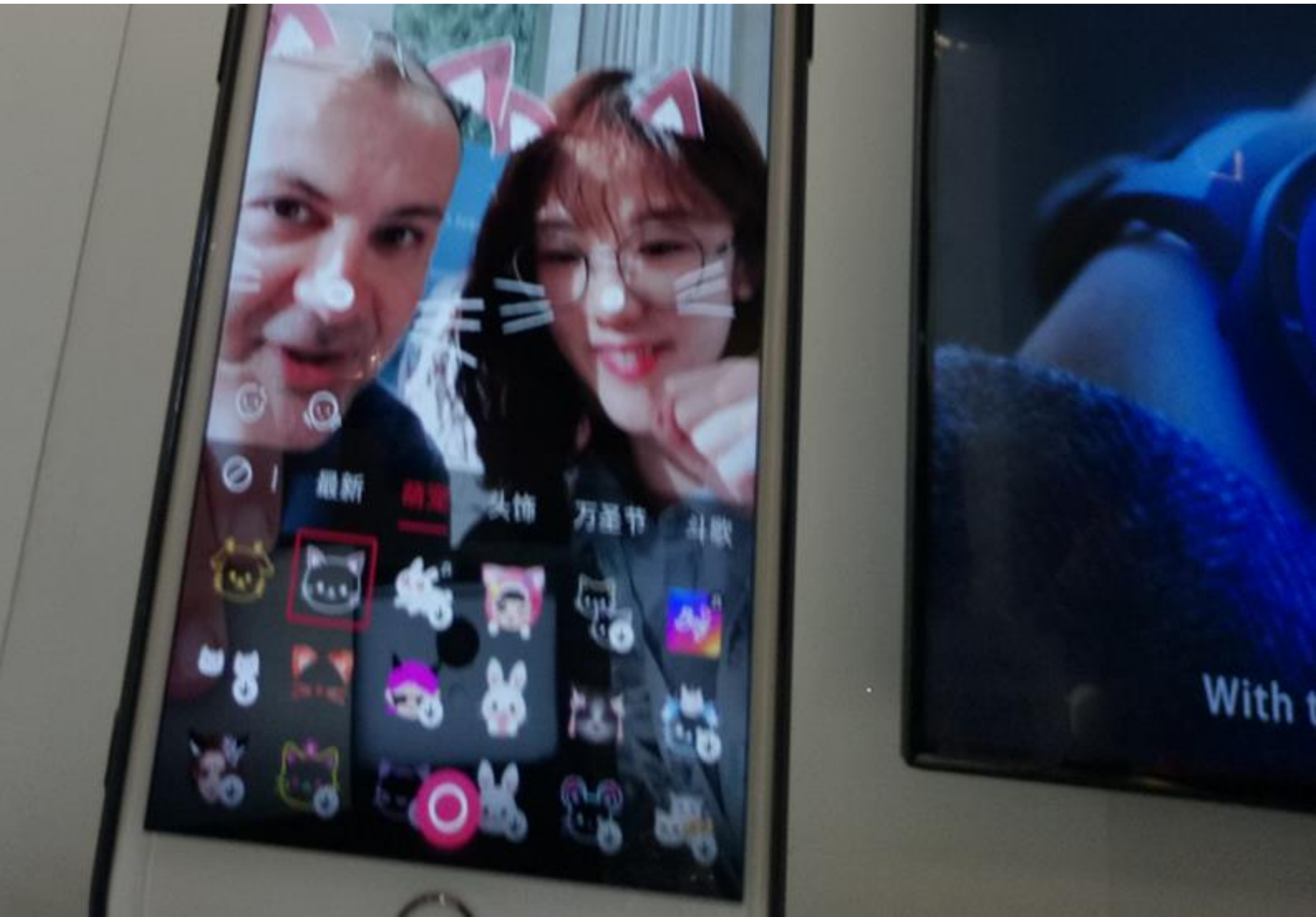
Mask R-CNN

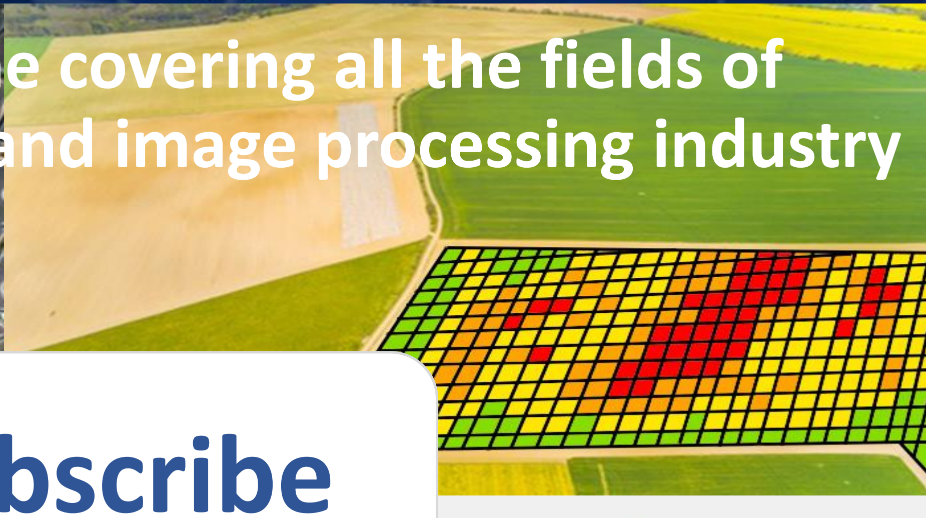
ICCV 2017, Venice, Italy

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick
Facebook AI Research (FAIR)

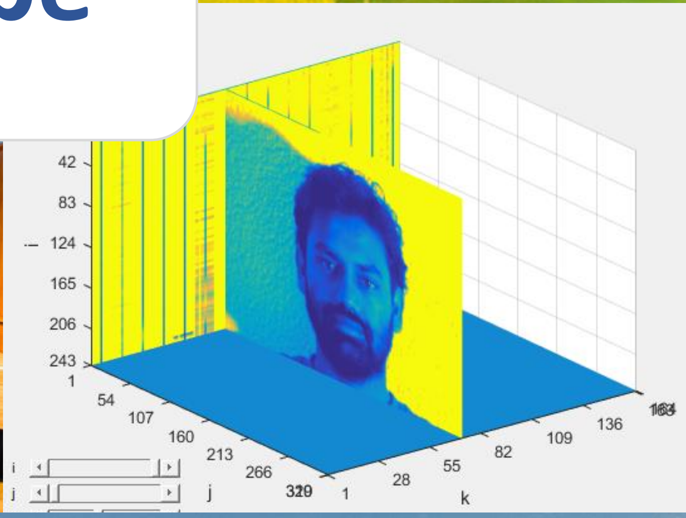
Sitting on stage: Cristian Smichisescu, Antonio Torralba

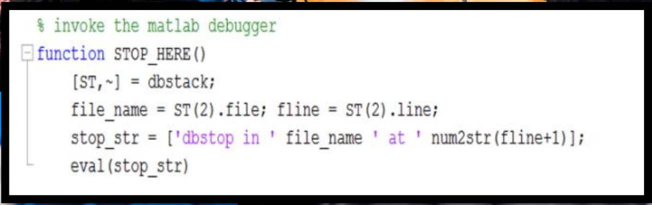
At the YouTu Lab booth at the ICCV 2017 Expo.





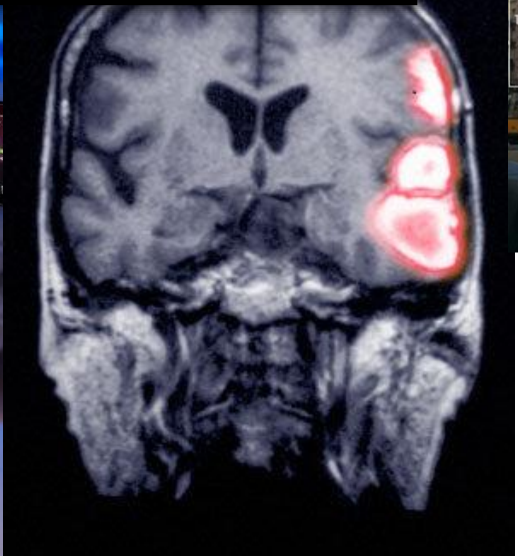
RE•WORK

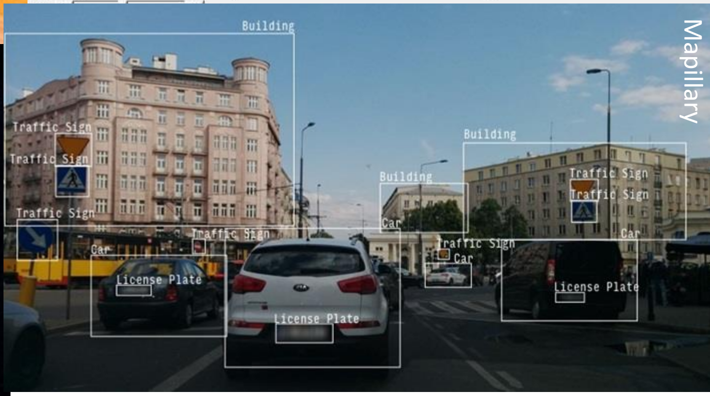




```
% invoke the matlab debugger
function STOP_HERE()
    [ST,~] = dbstack;
    file_name = ST(2).file; fline = ST(2).line;
    stop_str = ['dbstop in ' file_name ' at ' num2str(fline+1)];
    eval(stop_str)
```

Gauss Surgical





Mapillary