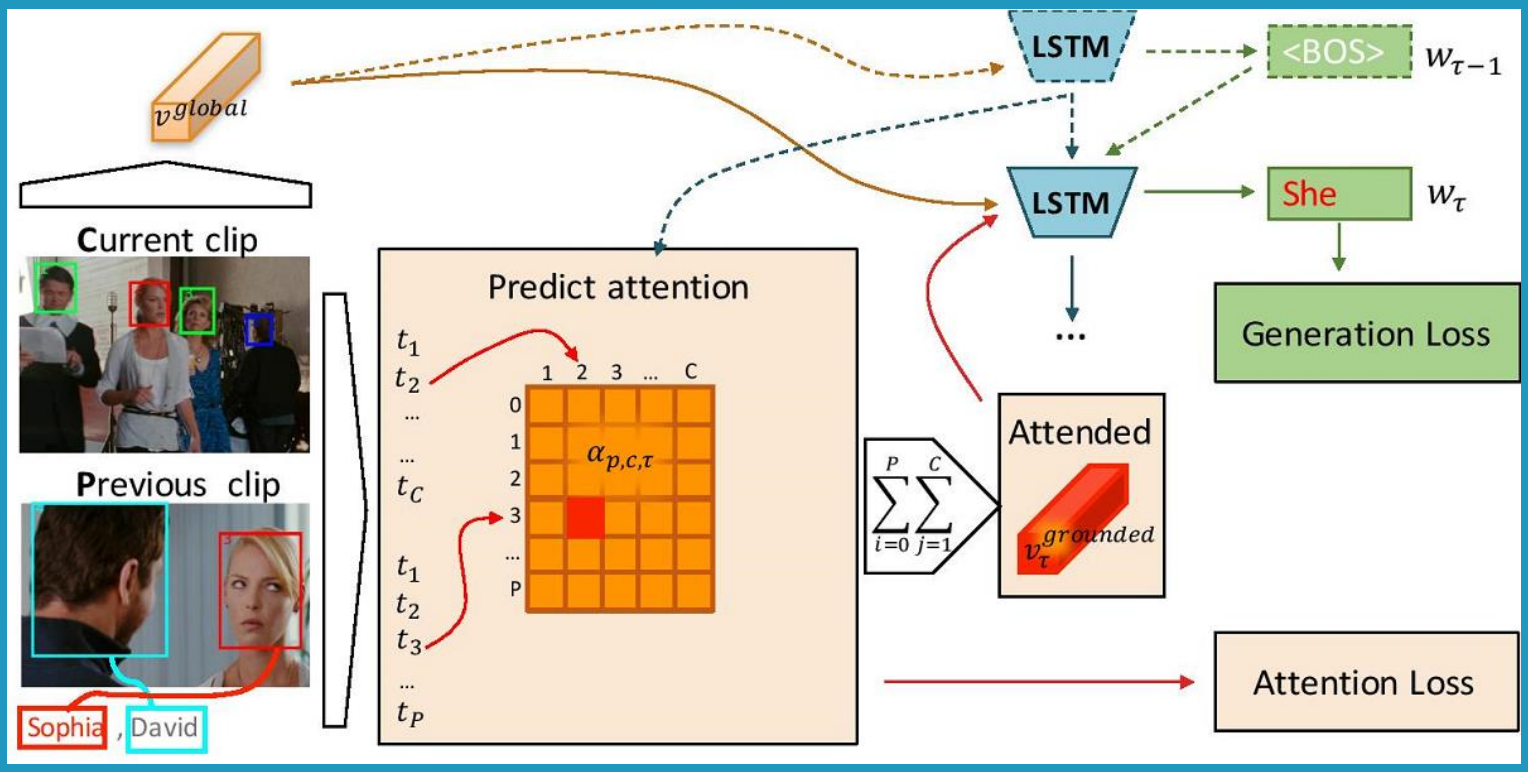
CVPR DAILY
July 21-26 HONOLULU 2017
Computer Vision & Pattern Recognition

Monday 24



Exclusive Interview with:
Nicu Sebe

Luca's Picks for today

Presenting work by:
Anna Khoreva
Anna Rohrbach
Namdar Homayounfar

Women in Science:
Amanda Song

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

A publication by
RSIP VISION

Luca's Picks

CVPR DAILY
July 21-26 2017 HONOLULU Monday

For today, Monday 24

Luca Rigazio

Luca Rigazio is Director of Engineering for Panasonic Silicon Valley Laboratory. Luca started one of the first Deep-Learning groups in the Valley, focusing on Vision and Learning for Robotics and Autonomous systems.

Don't miss Luca's picks for today, Monday 24!

- **Weakly/Zero-shot/Robotics/Relationships:**

P3-1.97 10:00 Page 30 of the Pocket Guide:
Weakly Supervised Affordance Detection
P3-1.90 10:00 Page 30 of the Pocket Guide:
Weakly Supervised Cascaded Convolutional Networks
S3-1A.18 08:50 Page 27 of the Pocket Guide:
Semantic Autoencoder for Zero-Shot Learning
O2-2A.9 13:33 Page 22 of the Pocket Guide:
Detecting Visual Relationships With Deep Relational Networks
S3-1B.15 08:38 Page 21 of the Pocket Guide:
Modeling Relationships in Referential Expressions With Compositional Modular Network

- **Theory**

O3-1A.12 09:45 Page 27 of the Pocket Guide:
Global Optimality in Neural Network Training

- **The required GAN paper**

S3-1A.8 10:00 Page 30 of the Pocket Guide:
Age Progression/Regression by Conditional Adversarial Autoencoder

- **Cool+Italian :)**

O3-1B.24 09:45 Page 28 of the Pocket Guide:
Semantic Autoencoder for Zero-Shot Learning

Luca's Picks

02

Women in Science Amanda Song

08

Anna Rohrbach

14

Nicu Sebe

04

Anna Khoreva

12

Namdar Homayounfar

16

Aloha, CVPR!

52,000 pageviews in less than 2 days for CRPR Daily… You probably enjoy our work and our content.

Thank you. Really. We are doing everything to make sure that you like also the next two CVPR Daily: the one that you are reading now as well as tomorrow's. Please share them with colleagues and friends!

Have a great day at CVPR 2017!

Ralph Anzarouth
Editor , **Computer Vision News**
Marketing Manager, **RSIP Vision**

CVPR Daily
Publisher: **RSIP Vision**
Copyright: **RSIP Vision**

Our editorial choices are fully independent from CVPR.

Nicu Sebe

CVPR DAILY
July 21-26 HONOLULU 2017 Monday

Nicu Sebe is the head of the Computer Science Department at the **University of Trento in Italy.**

We are here in Honolulu, Hawaii. It's already a very exceptional edition of CVPR with a record attendance of close to 5,000. What makes the concept of CVPR so dear to our community?

I think it's a trend over the last few years. One of the main reasons is that finally computer vision has had some very important success stories. More people have become interested in our problems. More people are realizing that what we do has an impact on the society. More importantly, the success of this conference is that industry has finally taken our ideas and uses them for commercial products. I think this is also happening in other computer vision top conferences. Our customers are finally realizing that what we do could be interesting to them. We do mostly software technologies, but also the hardware is catching up. We realize we cannot do our work unless we have good hardware. The time has finally arrived to do things together.

"One of the main reasons for the success of CVPR is that industry is meeting our students. This is really a market place."

What can we do to have a better blend between industry, academia, and the labs?

It depends on who is driving whom. In some sense, our students force us to work closely with industry. One of the main reasons for the success of CVPR is that industry is meeting our students. This is really a market place. Our students are asking us to model ourselves towards industry. I think going into that direction it's a good alternative.

You want to ensure that your students have a very successful career. At the same time, you may lose your most talented students because the industry has such a large budget compared to academia. How do you handle the will to have your students succeeds with the will to keep the best students?

"No matter what happens, I always win"

I always tell my students that no matter what happens, I always win. I agree with you that it's a pity to say that I'm losing my best student, but on the other hand, I'm not losing them. If they go into the industry and get a good position, I can always send my future students to do internships with my former students. So no matter what's happening, I think we as academia are going to win. I'm always sure that there will be good students, and there will be people that will want to be in academia. Not everybody is going for the money or for the big success. I'm not worried about losing our brains to academia. If we do it right, we as professors, we always win.

Some smaller companies complain that the big companies get all of the talents. The larger companies then complain that the startups offer more flexibility. Who do you think has a right to complain?

All of them have a right to complain, of course. All of them are right to some degree. It depends on the people. Being in a big organization has clear benefits. You have access to data, big infrastructure, and good colleagues. On the other hand you have some limitations. At a startup, you have flexibility. You have group dynamics and you work closely together with friends. There would always be both kinds of people working for the large companies or for the startups. They are two different worlds so I think people would like to try both. Some of my students have decided to go to startups on purpose even if they could have gone to one of the big companies. They want to try something new and take a chance. Maybe this startup is going to successful. Maybe it will be the next Google. This is how the big companies start. Big companies like Snapchat started like this. Many of our students are willing to dream.

"I don't believe in this kind of conspiracy theories…"

Some people are concerned that that there will be a shortage in STEM professionals. They say people will fight for the rare talent. Do you agree with that alarm or do you think people are exaggerating?

I don't agree. I don't think that there will be a lack of talent. The world is evolving all the time. I don't think that there will be a lack of talent or new ideas. The community is very successful. That means it will attract new, young people. This kind of shortage could only happen if something catastrophic happened. It's just a natural process. I think everything is going to evolve. I don't believe in this kind of conspiracy theories.

Nicu Sebe

So you don't think people should be alarmed? They will find the talent that they need?

Absolutely - I was talking to friends working for corporations saying how they compete with each other. I don't really look at it that way. Right now it is a fair competition. There are many resources. As long as you don't try to replicate each other, there is enough space for everyone in some way. How do you make sure you always get the best people? You have to offer the best conditions, not necessarily money. I mean good conditions in many ways like

"I believe we are finally in a golden era"

the perspective or the freedom to publish. People are choosing between these kind of things. It's a complete portfolio. I believe we are finally in a golden era. We have to benefit from that. We should work together. Competition is great. It helps to come up with better ideas and new ideas. So why don't we just do it together? I think that is the future.

Probably when you were a student, people would submit one or two papers each year to the big conferences. Now, it seems people must try to present a paper to every conference. How can a student find 6 or 7 great ideas every year?

It would be hypocritical to assume that we all have new ideas starting from scratch. For a student, the first paper is difficult. I always tell my students that their first publication will probably take a year or year and a half. Once you get there, you know exactly what is expected of you. Then you build on top of that. This is also a matter of success. If somebody has found a way to publish and present their work then they will manage to publish more. It's true that we might be over-publishing, but this is also a measure of success.

What was the best thing you ever learned from one of your students?

Probably endurance, I would say - If you fail a few times, they keep going. Of course, with a little bit of counseling, you tell them that next time they are going to be lucky, but I think it comes down to perseverance. I like students that have strong opinions, and when they believe in something, they don't give up. It can be confidence or even craziness up to a point, but I think it's good as long as it controlled. It's good to have crazy people in an inventive way.

"Multi-modal fusion is becoming part of the main trend"

Since you were a student, what technological development impressed you the most?

I would put it in a negative way. I think what impressed me is the fact that people, especially in computer vision, have not been using other modalities. Also the fact that up until now, computer vision has tried to solve the most difficult problems. Just assume that I can only see, but I don't hear. I don't smell. I don't do any other things. It's avoiding the fact that we are, as human beings, multi-modal systems. This is changing now. I can see it even now at the conference that many of the works are trying to use different modalities. Up until now, the community was considering only vision. What is now interesting and impressive is that now the multi-modal fusion is becoming part of the main trend.

Women in Science

Amanda Song is a third year PhD student working at **UCSD (University of California, San Diego)**. Her home department is cognitive science and her research topic lies in the intersection of computer vision, machine learning, and social science.

Amanda, can you tell us about your work?

I would define my work as computational social psychology which means I use machine learning technologies to study people's cognition and social cognition.

How did you decide to work in a field like this? What fascinates you about it?

There are a lot of interesting, random factors that led me to where I am. Initially, I started as a biology student. My first approach to science was about neuroscience and visual science. I studied the brain system, and how the visual cortex works at a single neural level. I dealt with animal experiment and single-unit recording. Later, I studied machine learning because I felt that the computational and machine learning model might help me have a better description and representation of what the human brain is doing. That's what I studied in the University of California San Diego's Cognitive Science Department. This department gives you a lot of freedom to do all sorts of research. You can take different approaches and combine them together in a creative way.

What kind of animals did you work with?

Cats and monkeys.

When you worked with animals, what did you find different and similar? Which of their features helped you understand the human brain?

At a single neuro level, I think that animals and humans share similar regional functions. For example, the primary visual cortex acts as the edge and contour detector. So in this aspect, it's very similar. The most distinct difference might be at higher level cognition such as language and logical reasoning. This is very different because a lot of human complex social behavior is based on our logical thinking and on our language to communicate our thoughts. I can't study that in animals easily.

You look at fields like language, vision, cognition, and so on. Science has taken a long time to try and understand how these things work together. What would you like to achieve? Do you think you can understand how the human brain functions at a higher level?

Well, that's a very big question. For me, the higher level question that I care about is to have a diverse, and yet complete

profile of human culture. For example, my specific research topic is humans' first impression on each other based on the visual input. For that, interestingly, humans share a lot of consensus. Although you may think first impressions are a subjective thing, people agree more or less on who looks more trustworthy, friendly, intelligent, and responsible. There are some common visual factors that drive our common consensus. Humans still have slight differences on how they perceive the world and how they perceive each other. I would like to know how demographic and personal experiences drive those individual differences. I would like to have an individual model for everyone in how they understand the world.

Does your work influence you in such a way that it changes how you look at other people? Do you try to understand how they react to things?

[laughs] Yes - In research, we just fill the dataset. We show people images of people, and then we ask them for their subjective first impression of those people. In this way, you can collect people's responses, and you can model people's average response as a human population average perception about a person. You learn the mapping of the image of the person's average impression of that photo.

Do you observe how you react when you meet a person for the first time?

[laughs] That's interesting! For the first impression, there's a potential bias in our perception. For example, people may associate males with more leadership or females with family... these kind of associations. Similarly, we can observe some sort of trend in your first impressions. You may feel like a Caucasian person looks more trustworthy compared to an African American. Some people may have those associations. If we are aware of our implicit bias, we might be able to fight against it and be more rational.

Did you see any difference between genders in their reactions?

This is a very important question. The first factor you may think about is the gender difference. For now, we are using a public dataset collected by a MIT group. It's not our own dataset. In this dataset, we don't have full access to the raters, the perceivers, or demographic information including gender. From this dataset, we are unable to answer this question. In the future, if we are going to build our own dataset, we will collect every raters on demographic information. Then we can answer that.

Your impression is that you expect to find some differences?

Maybe! At least for facial attractiveness, in general females are perceived as more attractive than males so… sorry about that guys [laughs].

Don't be sorry! We are very happy that females are more attractive than males!

[*we both laugh*]

Someone asked me recently: "Why should we even take gender into consideration? We should judge other people as a human being and for what they do. We should forget about gender." Do you think it is possible for people to completely ignore the gender of the person in front of them?

I think that would be very hard because it is our biological nature to quickly make a lot of inferences about the other person's gender, age, or ethnicity. It's a program embedded in our brain. Even if you don't want to do that, your brain may still subconsciously do that for you.

Your rational brain will try to ignore that or try to delete the bias. Maybe females are more suitable for certain jobs. You can try to delete the bias from you brain.

How did it become embedded in the human brain?

This is a very interesting question. You may find some evolutionary answers. For example, if you narrow it down, a lot of first impressions come to two key factors. One factor is about the intentions of the other person such as friendliness or trustworthiness. They all try to judge the other person's intentions. The other dimension is the other person's capacity to carry out the intentions. The first thing you need to know is if the other person is a friend or an enemy. If he's a friend, will he be able to help me? If he's an enemy, will he be able to harm me?

There are two dimensions: intention and capacity to carry out the intention. You need to quickly judge that in order to survive in the ancient world.

Does this study suggest the changes people should make to their behavior?

Our future studies will try to answer the questions like to what extent do people associate, let's say, females, senior people, certain ethnic groups, etc with competence, confidence, or trustworthiness. If we observe people having implicit bias, we will reveal it, and call for actions to fight against it. It remains to be seen.

Does it change how you look at people and their behaviors? Do your studies affect your judgement

when you meet other people?

[smiles] Sometimes… When I realized that we are not immune to those first impressions, I was a little bit empathetic that this is our human nature. This is how we act and behave. This is how we are navigated by our first impressions. How sad is this? That's how I feel sometimes.

Your first reaction would be to see it as an expression of the virtues of the human spirit or the contrary?

I would say the contrary.

[we both laugh]

It's not that bad. It's just natural because we have to make a lot quick inferences in order for us to succeed or just to survive in the world. Those inferences are not accurate sometimes. Even if they are accurate, we shouldn't base our behavior on that. It's hard to overcome. It's hard to overwrite.

Would you recommend people to change their behavior if they see that it brings them to the wrong conclusions?

Yes

How would they be able to judge when they are having the wrong reaction?

There are two sides of the coin. Everyone want to be liked, look trustworthy, friendly, or competent. We all want to present the best of ourselves to other people. On the other side, people will judge you based on your appearance. Sometimes we don't want to be judged by our appearance, but we want to be judged by our behavior and for the person that we really are.

What are your thoughts about research in computer vision?

Right now in computer vision, you have a lot of exciting opportunities to do research in industry. I think there is a new trend for a lot of students to do internships during the summer or even during school years. They apply what they learn into a job in industry. Then they turn their research ideas into a product. I think this is very exciting and admirable.

Also deep learning has a lot of potential, especially when there's big data. It will change a lot of regions wherever big data is available, but it's not the only story. We still need to combine a lot of inference and traditional methods or other methods in order to complete the story.

CVPR DAILY July 21-26 2017 HONOLULU Monday

"Use static images to train the network"

Anna Khoreva is originally from Russia and is now doing a PhD at the Max Planck Institute for Informatics in Germany. Yesterday (Sunday) she presenter her paper here at CVPR, "**Learning Video Object Segmentation From Static Images**".

This is joint work together with **Federico Perazzi** from **Disney** and **ETH Zurich**, **Rodrigo Benenson** from **Google**, her supervisor **Bernt Schiele** from **Max Planck** and **Alexander Sorkine-Hornung** who was at Disney at the time of this work and is now at **Oculus**.

"In the end we do online training, and we fine-tuned the model on the first frame to learn the appearance of the object"

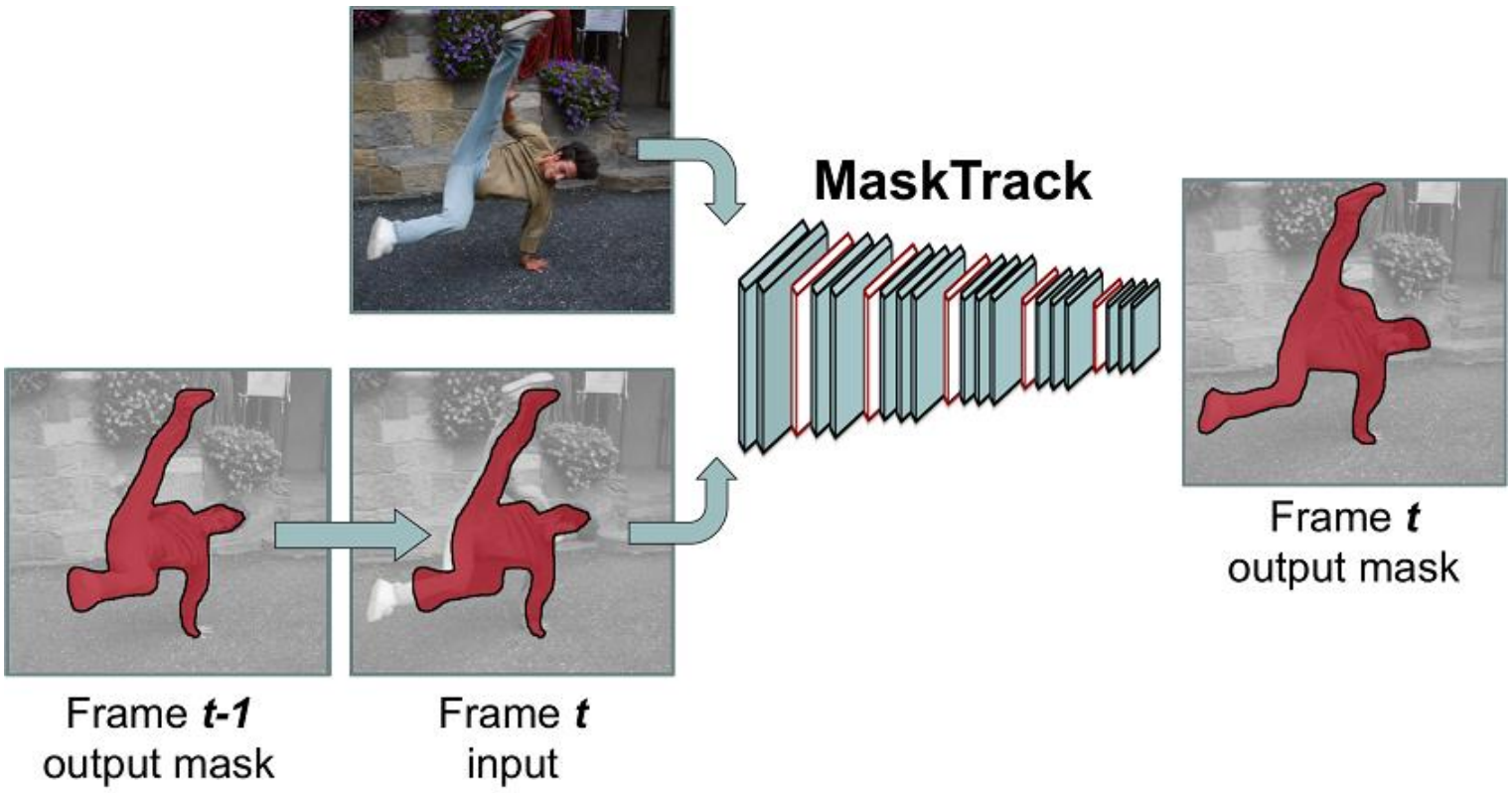
Anna told us about the motivation behind using static images to produce video segmentation: when training convolutional neural networks for video segmentation, you usually need densely annotated video data. However, there is a lack of such densely annotated datasets because it is necessary to have pixel-level mask annotations, and this is very expensive to annotate. Therefore, Anna and her co-authors propose to use static images to train the network. Such images are better available, and there are large scale datasets like PASCAL or MS COCO. Specifically, in this work they propose to treat video object segmentation as guidance instance segmentation, and to use the previous frame to guide the segmentation. They approach this per frame, which makes it very efficient and needs only one pass over the video to produce results. The training of the model involves different stages. They first do offline learning, and "*in the end we do online*

training, and we fine-tuned the model on the first frame to learn the appearance of the object". They used a fully convolutional neural network which was originally used for semantic segmentation, and re-purposed it to do binary classification to segment foreground and background. Additionally, along the RGB image, they also added an extra channel as an input with the binary mask from the previous frame output, Anna explained.

Anna says that "*in this work, we still use images with mask annotations for training, which is still quite expensive and takes a lot of human effort to annotate*". In a follow-up work, they propose to use synthetic data. They propose training the network using synthetic data which is generated by using the first frame and its mask annotation. The title of the paper is "Lucid Data Dreaming for Object Tracking", and will be presented at a workshop on Wednesday here at CVPR.

Related to this is also the DAVIS challenge on video object segmentation LINK and the related workshop about this challenge on Wednesday, which was co-organised by one of the authors, Federico Perazzi.

If you want to learn more about Anna's work, make sure to visit her follow-up work "**Lucid Data Dreaming for Object Tracking**" in the "DAVIS Challenge on Video Object Segmentation 2017" workshop, which will be held on Wednesday July 26.



CVPR DAILY July 21-26 2017 HONOLULU Monday

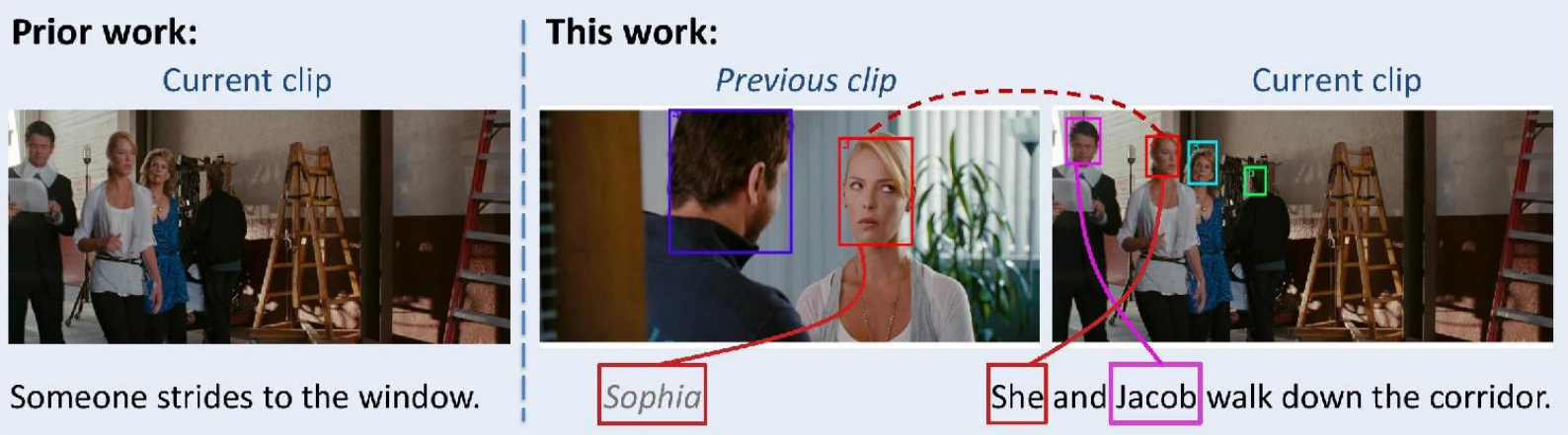
Generating Descriptions With Grounded and CoReferenced People

Anna Rohrbach did her PhD at the **Max Planck Institute for Informatics**, and the focus of her research is video description with natural language.

While prior work focused on describing a given video by producing a sentence, in Anna's current work, they are trying to extend this to a more rich and more interesting predictions. They want to answer questions like: What are the people wearing in the scene? What are their genders? Have we seen them previously? Where exactly are they? I.e., they want to localise objects and resolve visual co-references.

"*We are tackling a very complex problem*", Anna says, "*we describe the video with richer person-specific labels like gender, and we also localise them*". The advantage of this is that it allows them on the one hand to get a visualisation of what the model is doing, and also to inspect the errors which the model makes and understand what is going on in the video. The architecture they used for the model is very complex and includes many steps. They first need to detect people in movies with different view-angles and conditions - which is quite challenging on its own already. They also track people in the video and on top of this, they have to learn to associate the names with the visual appearances. "*And finally, we come to the actual problem we are trying to address*", Anna explained, "*where we have to do this description along with all this meta-information.*"

The most challenging thing for her and



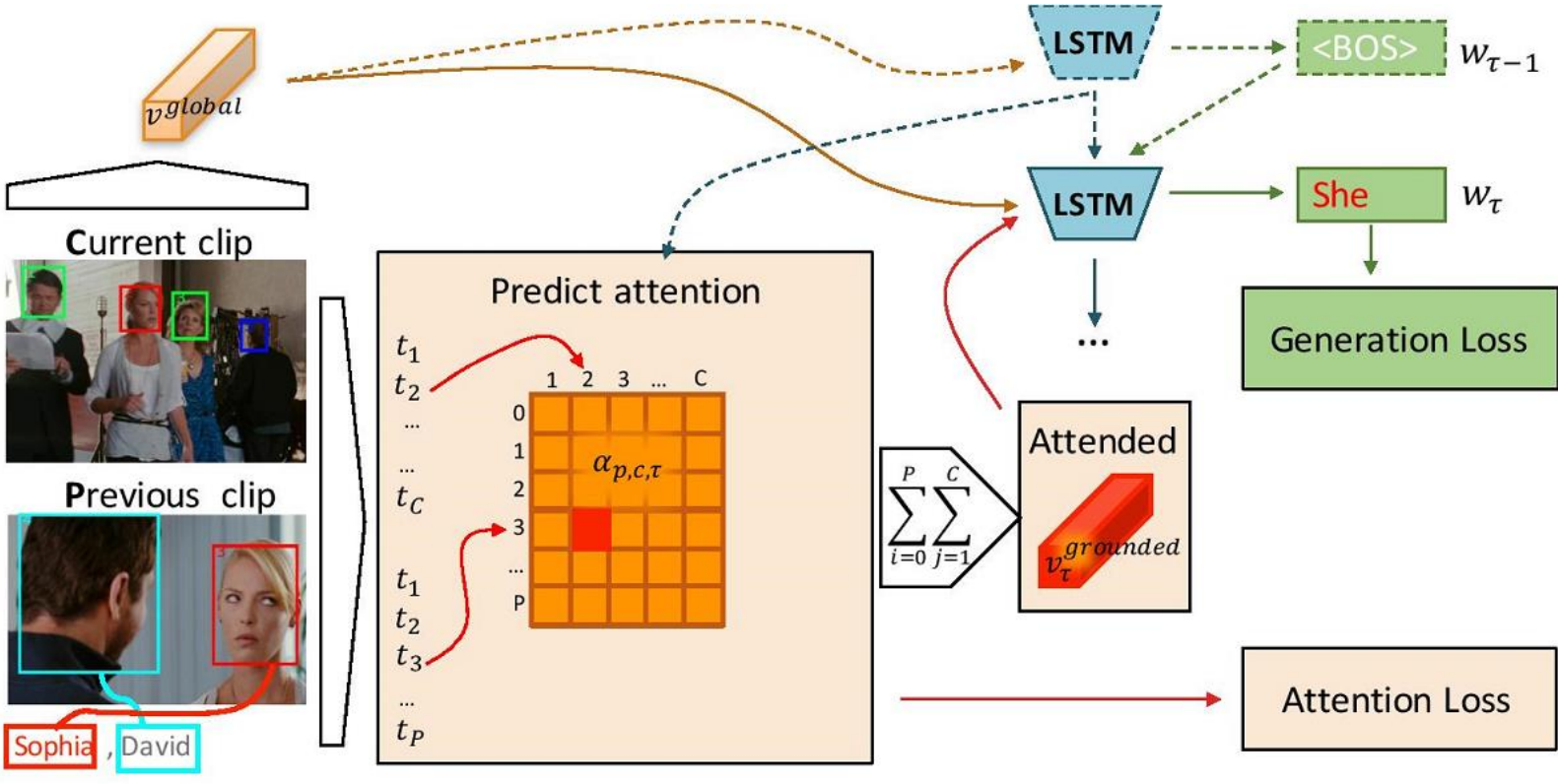
"Shoot for the moon; even if you miss, you'll land among the stars!"

her co-authors was to make the model learn this attention about the tracks with they have extracted in the video, and simultaneously try to describe the sentence correctly while predicting the additional modalities they need for their model.

To realise this, they used an **encoder-decoder LSTM** based approach. The core of their work is the attention mechanism which reasons about the tracks of the current clip and the previous clip, and which then jointly addresses the grounding and the co-reference challenges. This then informs the LSTM to predict the right person-specific labels.

Anna's most ambitious is that one day she would like to tackle the entire movie, and not only look into one clip back. She says: "*I would like to describe the entire movie consistently and coherently, and we are doing baby steps in this direction*". Talking about her supervisor **Bernt Schiele** (who is also involved in this work), she told that he always says: "*Shoot for the moon; even if you miss, you'll land among the stars!*". In this spirit, Anna is aiming at something very ambitious, and although they might not get there immediately she believes that they will get to "something cool".

"I am motivated by the idea of helping the visually impaired and blind people", Anna says, "so I hope that one day we will be able to automatically describe movies and other visual sources to assist them".



Sports Field Localization via Deep Structured Models

Namdar Homayounfar is a PhD student at the department of Statistical Sciences at the **University of Toronto** and is also currently interning at **UBER**.

The goal of their work, Namdar told us, is to localise a sports field in an image. Doing this is the first step in data generation in sports, because based on knowing where the field is, the players can be localised and more statistics about them can be extracted - like how much they run, which positions they occupy or to identify offsides.

The novelty of this work is the way they formulated and solved the problem: they are the first to use single-image (monocular) inputs, and their approach is fully automatic, very fast and exact.

Originally, Namdar was trying to solve a different problem - to automatically generate captions and statistics about the players. But soon he realised that in order to do so, they first need to localise the field. He tried using existing methods, but after a few months had to conclude that they did not perform well enough. "*This problem could be solved if we had four points from the image and four points from the model, and we tried to estimate the homography matrix directly*", he explains. But figuring out which four points in the image correspond to which four points in the model turns out to be a very difficult problem.

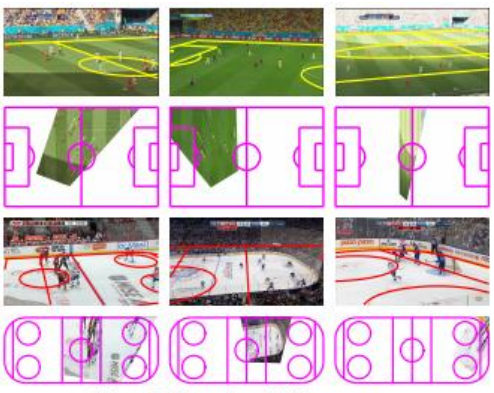
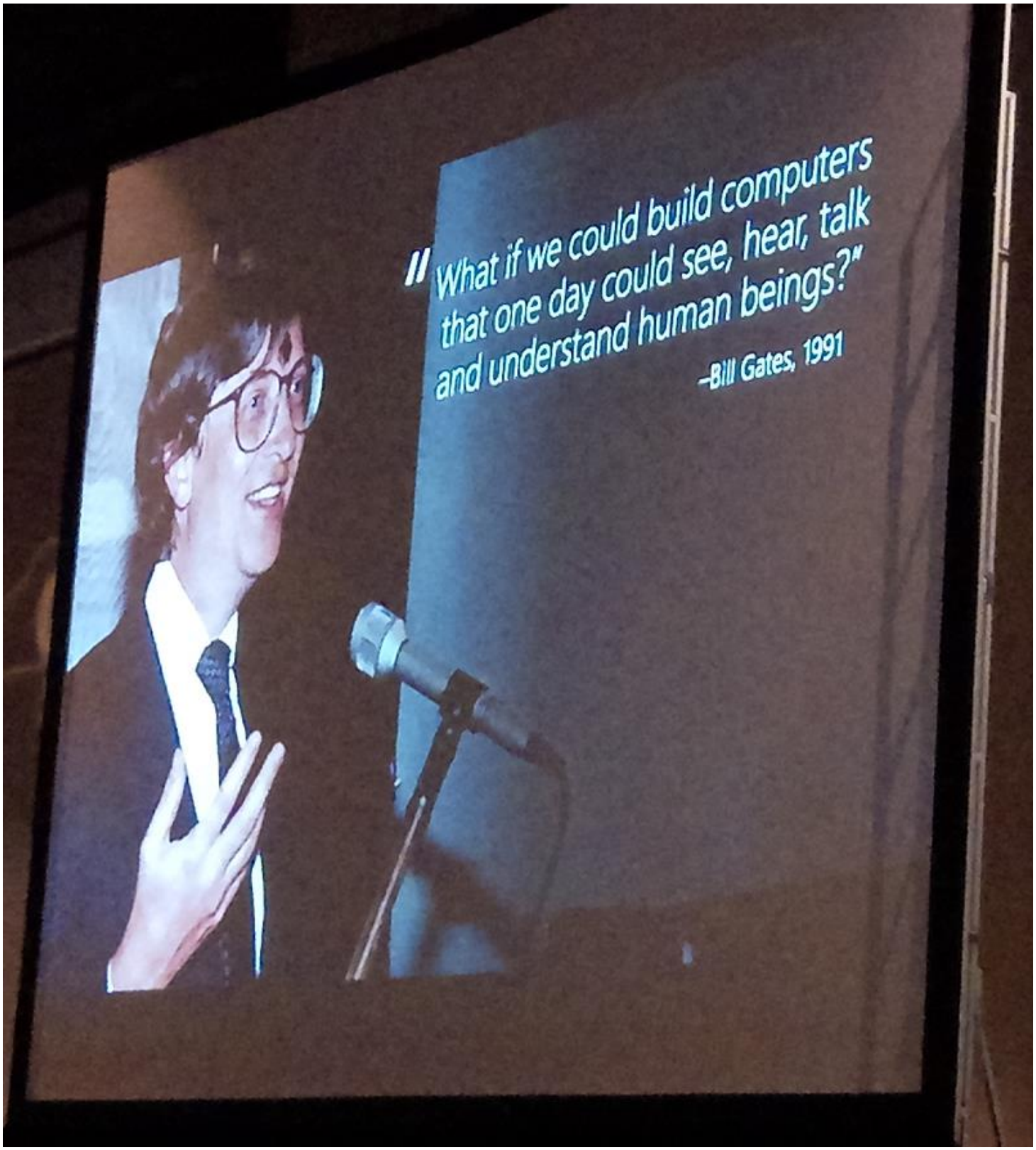


Figure 8: Examples of failure cases

Besides just localising the sports field, they wanted to also have additional information about it like where the grass is, where the lines of the field are, or where the outside of the field is. It is very hard to do this using handmade heuristics, due to the variations between different fields. Therefore, they came up with a new machine learning method to solve all of these problems of field localisation.

They use a neural network model that is able to tell them per pixel what exactly it contains, and thus solves the problem of both field localisation and answering more specific questions. This predictions are done per image, and Namdar tells us that in the future, they want to do this in a temporal manner, for videos instead of single images, incorporating temporal priors so that there is a smooth transition of homographies between the frames.

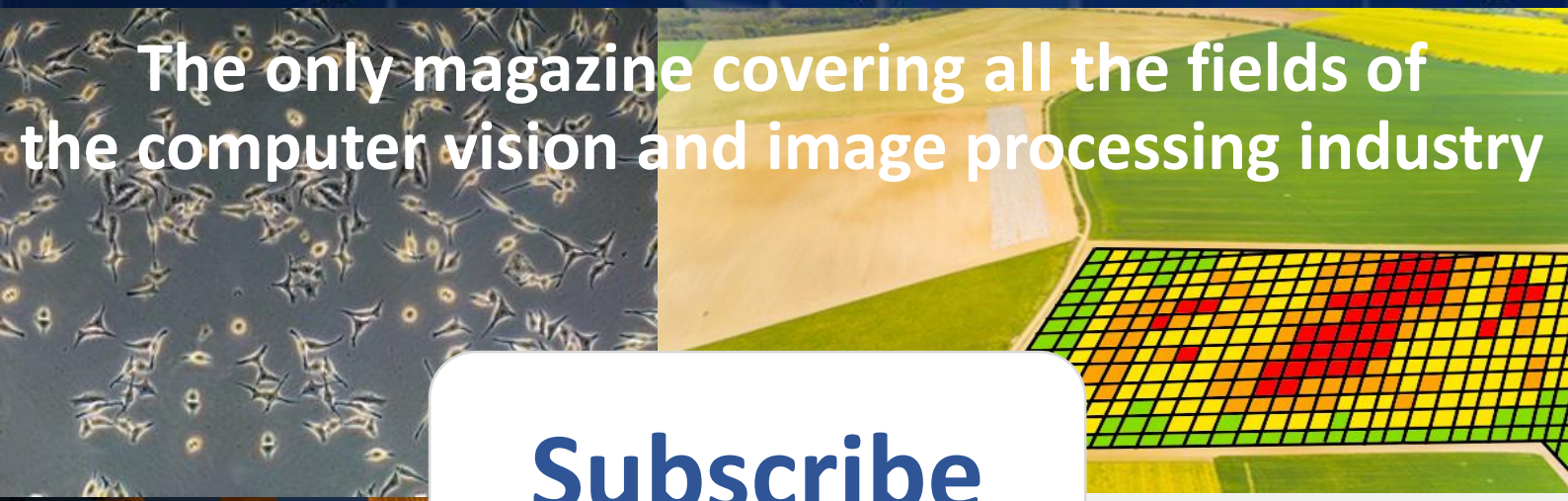
Namdar's supervisors and co-authors, **Raquel Urtasun** and **Sanja Fidler** followed attentively our discussion; when asked to mention the main attractiveness of this work, Raquel told us that "*the key of this work is to come up with a parameterisation of the problem that allows you to do efficient inference by taking into account the structure of the problem and the advantages of convolutional neural networks and deep learning*".


"What if we could build computers that one day could see, hear, talk and understand human beings?"
–Bill Gates, 1991

Improve your vision with

Computer Vision News

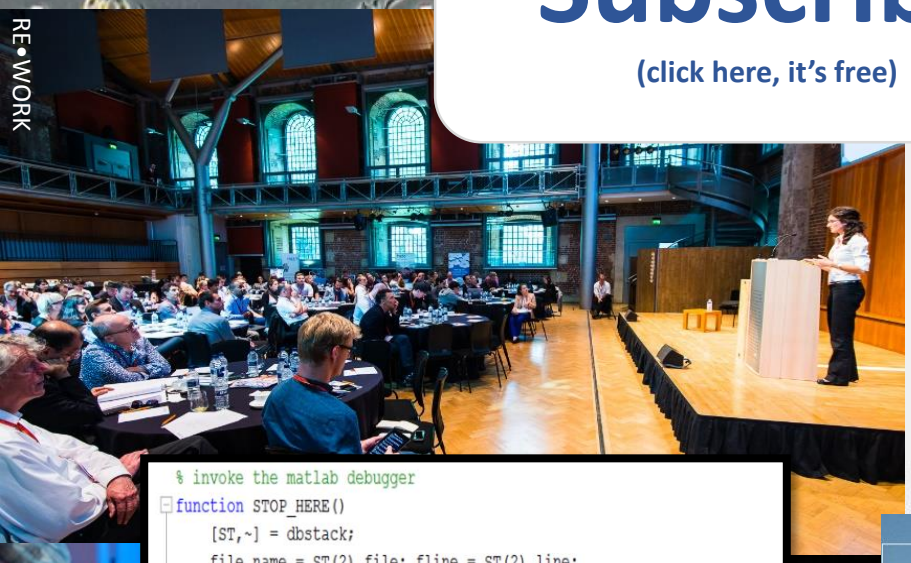
The Magazine Of The Algorithm Community

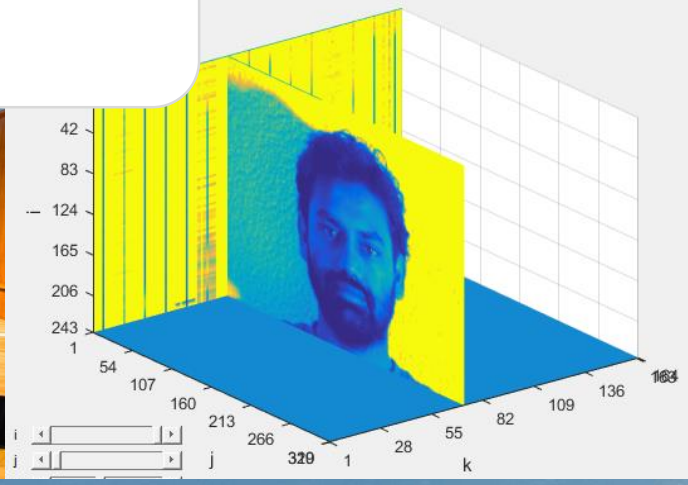


The only magazine covering all the fields of the computer vision and image processing industry

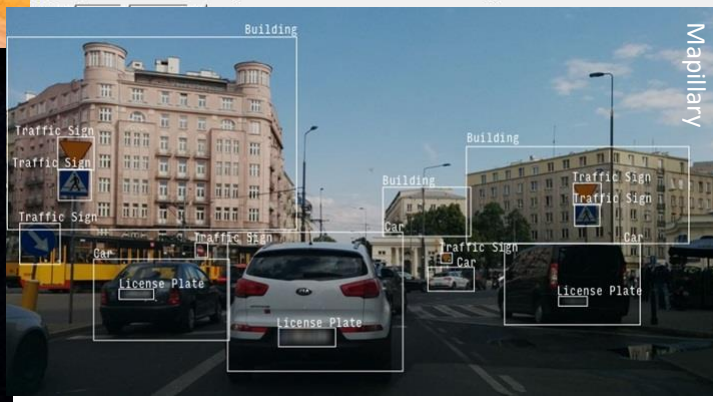
Subscribe

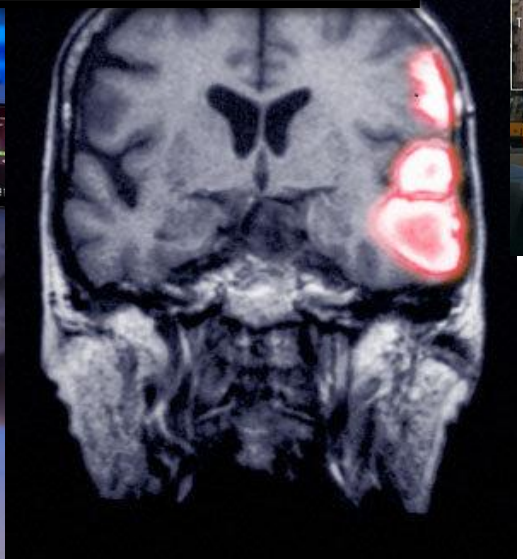
(click here, it's free)






```
% 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)
```





A publication by

RSIP VISION