Computer Vision News

The magazine of the algorithm community

August 2017

Exclusive Reviews by RSIP Vision Engineers:
TransFlow: Unsupervised Motion Flow...
GMS: Grid-based Motion Statistics...

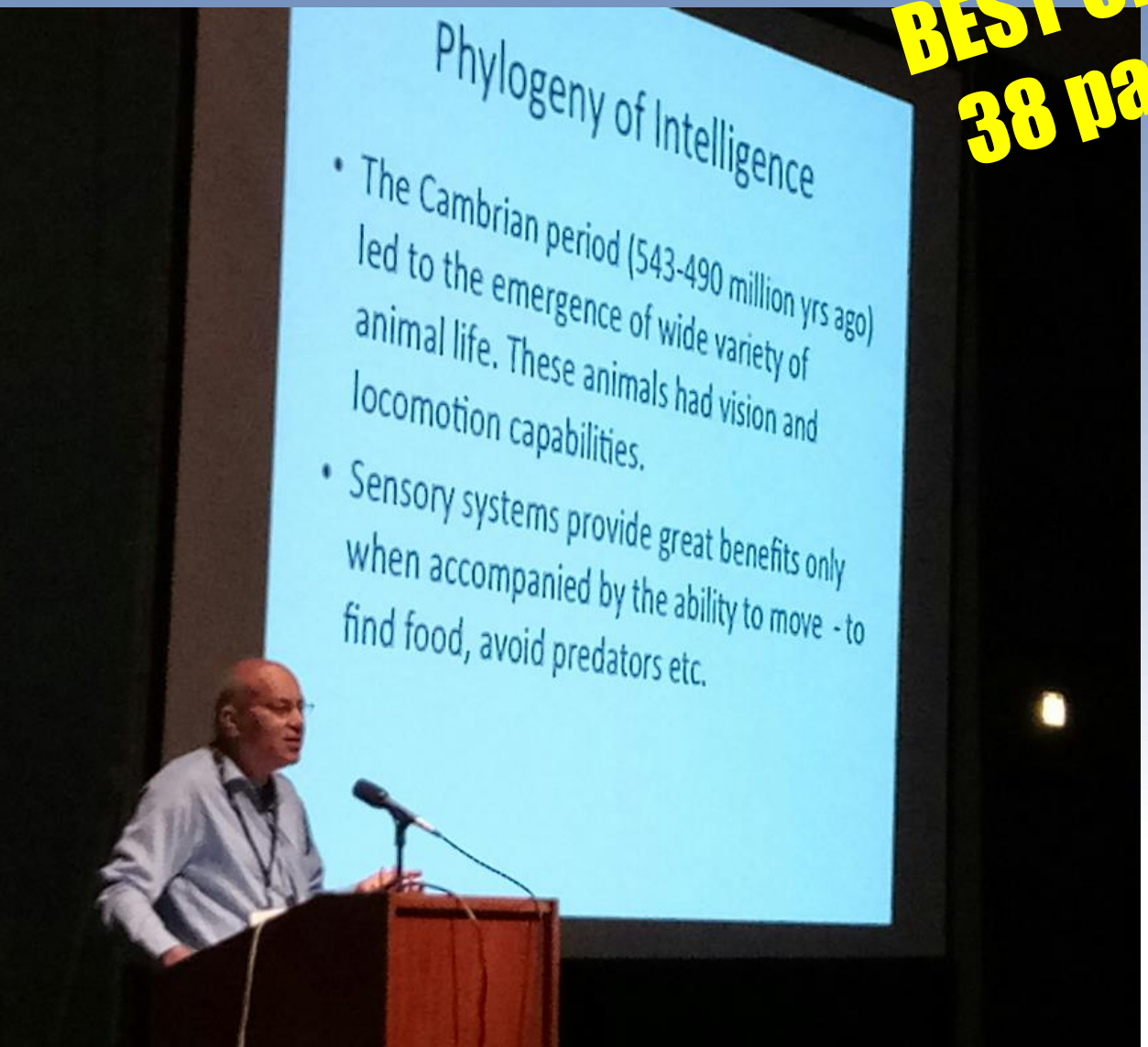
Women in Science:
Amanda Song
and
Nour Karessli

Workshop and Challenge at CVPR2017:
Densely Annotated Video Object Segmentation

Presentations at CVPR by:
Linjie Li
Anna Rohrbach
Angela Dai
Derya Akkaynak
Phillip Isola
Namdar Homayounfar

Exclusive CVPR Interviews:
Sanja Fidler
Vittorio Ferrari

BEST OF CVPR:
38 pages !!!

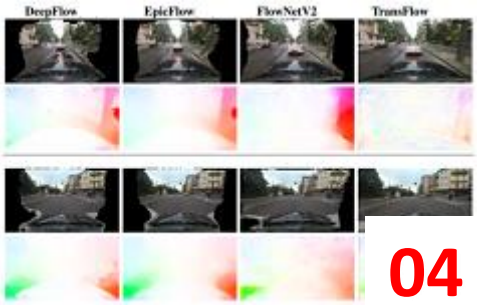


Spotlight News

Computer Vision Events

A publication by
RSIP VISION

Research
TransFlow by Panasonic



04

Phillip Isola

25

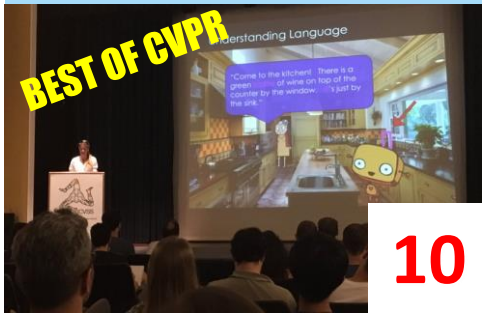
Challenge: DAVIS

44

Research
Grid-based Motion Statistics

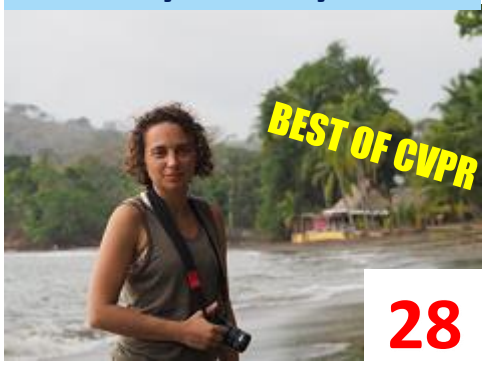
48

Sanja Fidler



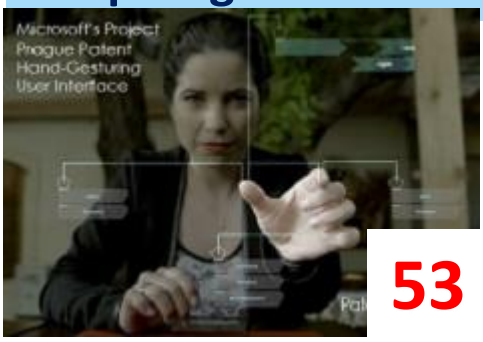
10

Derya Akkaynak



28

Spotlight News



53

Vittorio Ferrari

20

Women in Science
Amanda Nour

30 38

Upcoming Events

55

Our top reporters from CVPR2017 in Honolulu, Hawaii (Image courtesy of Erika Roberts)

Computer Vision News
Editor: **Ralph Anzarouth**
Publisher: **RSIP Vision**

Contact us

Give us feedback

Free subscription

Read previous magazines

Dear reader,

CVPR2107, the Computer Vision and Pattern Recognition conference, has once again fulfilled the expectations of our community: an exceptional technical program combined with a dream location in Hawaii and an outstanding organization produced the most important meeting of the year. A record-breaking **5,000 participants** enjoyed a perfect week and, just like last year in Las Vegas, **Computer Vision News** was a key actor in the success of the event: our **CVPR Daily**, the magazine offered by RSIP Vision to the algorithm community during CVPR, was a success renewed every day among all attendees. If you missed it, you will find in this August issue of Computer Vision News a **Best of CVPR** section with no less than 38 pages, all relating some of the key moments that we lived in Honolulu. We are very proud of this renewed partnership with **CVPR**, and have already agreed to repeat its success at **CVPR2018** in **Salt Lake City, Utah**. We wish to warmly thank **Nicole Finn and CtoCevents** for their precious help and support.

The success of CVPR Daily does not conceal from our eyes the exceptional quality of the **technical program** and the **impressive talent** of all the scientists who presented it. A quick look at the program of the conference would convince anybody to join us next year for an even more successful CVPR conference.

This Computer Vision News issue of August offers many other exclusive reads, in particular our own reviews of two great research papers: **GMS (Grid-based Motion Statistics)** and **TransFlow** (sponsored by **Panasonic**).

Please keep sharing this magazine and **enjoy the reading!**

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

TransFlow: Unsupervised Motion Flow...

Every month, Computer Vision News reviews a research from our field. This month we have chosen to review two papers. The first one is **TransFlow: Unsupervised Motion Flow by Joint Geometric and Pixel-level Estimation**. We are indebted to the authors (**Stefano Alletto, Davide Abati, Simone Calderara, Rita Cucchiara, Luca Rigazio**) for allowing us to use their images to illustrate this review. Their work is here.

**Panasonic Silicon Valley Laboratory sponsored this work in cooperation with Unimore, the University of Modena and Reggio Emilia.
Imagelab Unimore is grateful to Panasonic for its generous sponsorship.**

Background, motivation and novelty:

In the last few years, there is a growing interest among computer vision and machine learning researchers about **autonomous and assisted driving**. Optical flow estimation is one of the most researched topics in this context, but remains an open problem. The automotive context: large displacements, extreme changes in lighting conditions and the automotive movement making objects' unique motion patterns very difficult to disentangle -- make optical flow particularly challenging. Moreover, solutions that use deep learning typically require large annotated datasets, however pixel-level annotated datasets are lacking in the automotive field.

TransFlow builds upon several previous deep learning methods: **the Spatial Transformer (ST)**, developed by the DeepMind team in 2015, based on convolutional neural networks, this unit can be added to any network to perform explicit spatial transformations of features. Spatial transformers produce models that learn invariance to translation, scale, rotation and other generic warping. **FlowNet**, developed by Fischer et al, is one of the first end-to-end deep architectures for dense optical flow. Using a convolutional-deconvolutional autoencoder FlowNet provide a solution to the problem posed by the absence of large annotated datasets by synthesizing an image dataset featuring random chairs flying over random landscapes. The information is first spatially compressed in a convolutional encoder block and then refined and re-expanded in a transposed-convolutional decoder block, in a mirror architecture.

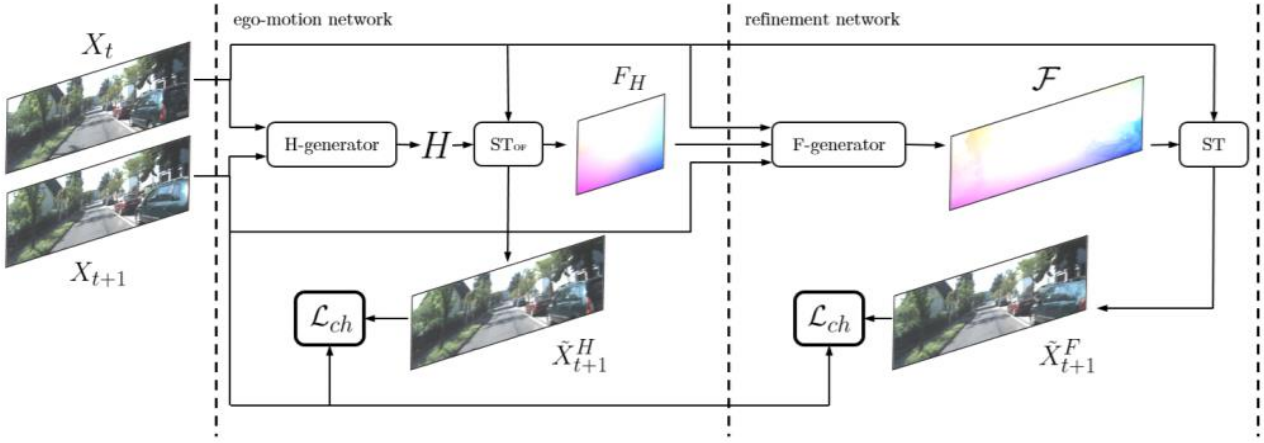
The novelty and advantages of TransFlow, compared to existing studies conducted so far, are: (a) Significantly better generalization capabilities compared to supervised approaches; (b) a simple and fast solution; (c) an end-to-end forward-only neural network; and (d) outperformed other recent attempts at unsupervised optical flow estimation.

Method:

The input and output of the TransFlow method:

	Training	Testing
Input	A pair of successive input frames denoted by X_t and X_{t+1}	Input frame X_t 1. H-generator transform 2. ST network weights 3. F-generator network weights
Output	1. Homography transform 2. ST network weights 3. F-generator network weights	Estimate frame $\widehat{X}_{t+1}$

The architecture of the method is illustrated in the following figure:



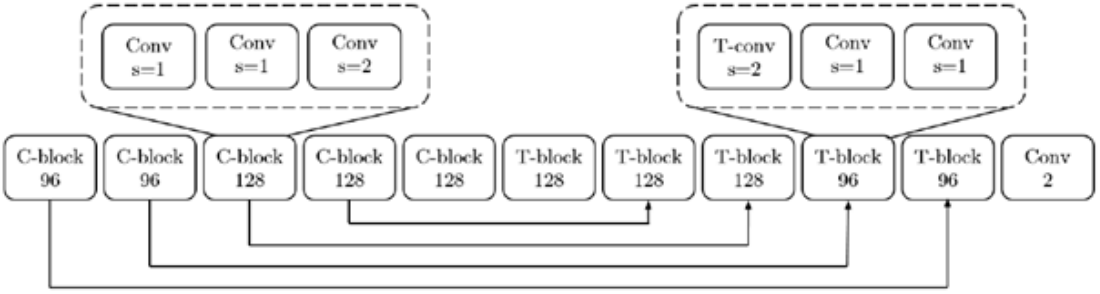
TransFlow consists of three main steps: (1) ego motion estimation; (2) motion refinement; and (3) edge aware smoothing:

1. Ego-motion estimation is a global flow step approximating the motion of the car. It consists of two parts:
   a. H-generator network produces a dense global flow of the overall motion
   b. ST - spatial transformer layer warps-in on main content.
2. Motion refinement - produces a fine-grained flow. It also consists of two parts:
   a. F-generator deeper network (structured similar to FlowNet) produces dense pixel-level transformation.
   b. ST - spatial transformer layer warps-in on main content.
3. Edge aware smoothing - aims at uniform flows within object boundaries.

Next is a detailed description of each of the 3 components of the architecture:

(1) The layers of the H-generator network are illustrated below, the H-generator produces a global flow representing the motion, without the details of individual objects. During training the H-generator concatenates the two frames Xt and Xt+1 and outputs the 9 parameters of the homographic matrix transformation.

(2) The F-generator network layers are illustrated below. The F-generator learns to refine the flow by taking account of moving objects and fine details ignored by the H-generator network and at the same time keep a consistent flow.



The F-generator network has a mirror structure: 5 convolutional layers that form the encoding block, followed by 5 transposed-convolutional layers that form a decoder block. Each of the 10 layers includes three 3X3 convolutional layers with leaky ReLU activations. A final top 2-channel convolution layer produces the final optical flow in the range of -1 to 1.

The global loss function employed during the training is a weighting the errors of the H-generator and the F-generator networks.

$$\mathcal{L}_{ch} = \mathcal{L}_{ch}(\tilde{X}^{H}_{t+1}, X_{t+1}) \times \alpha + \mathcal{L}_{ch}(\tilde{X}^{F}_{t+1}, X_{t+1}) \times \beta$$

Where $L_{ch}(\hat{X}_{t+1}, X_{t+1})$ is $\sqrt{(\hat{X}_{t+1} - X_{t+1})^2 + \varepsilon}$

(3) Edge aware smoothing- Edge aware smoothing improves the optical flow estimation as it allows to get uniform flows within object boundaries, which

often correspond to motion boundaries. Its output of a high dimensional Gaussian filter is given by:

$$\tilde{z}_i = \frac{\sum_{j=1}^{N} z_j K(f_i, f_j)}{\sum_{j=1}^{N} K(f_i, f_j)}, \quad \text{with} \quad K(f_i, f_j) = e^{-(||f_i - f_j||^2)}$$

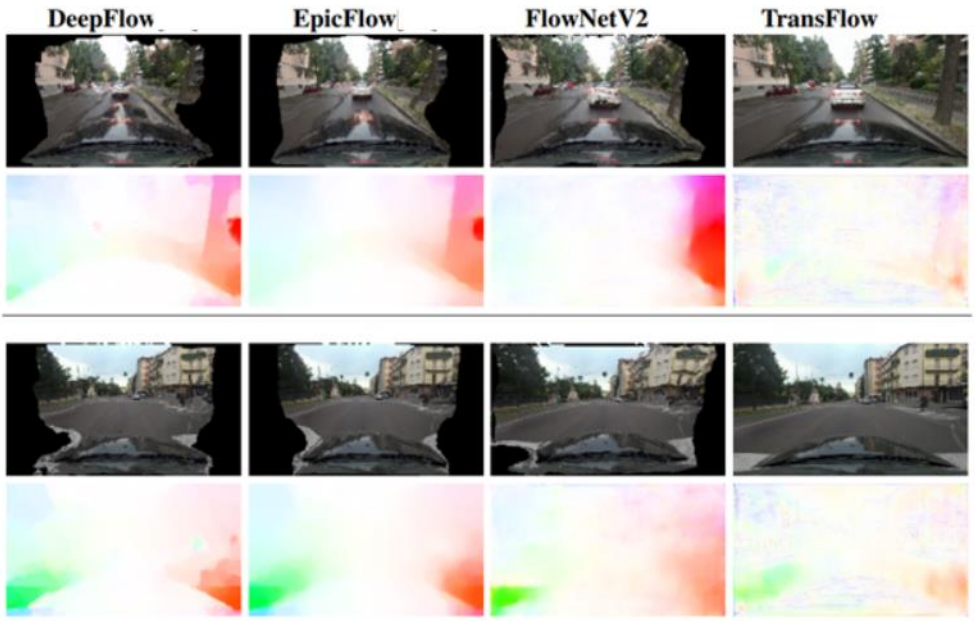
where

$$f_i = \left(\frac{x_i}{\sigma_s}, \frac{y_i}{\sigma_s}, \frac{r_i}{\sigma_c}, \frac{g_i}{\sigma_c}, \frac{b_i}{\sigma_c}\right)$$

in which x_i and y_i represent the pixel's location within the image; r_i , g_i and b_i encode the color of the pixel in the target image and σ_s and σ_c are hyperparameters tuning the sparseness of spatial and color features.

Evaluation and Results:

Qualitative assessment of frame reconstruction from the KITTI Flow 2012 dataset:



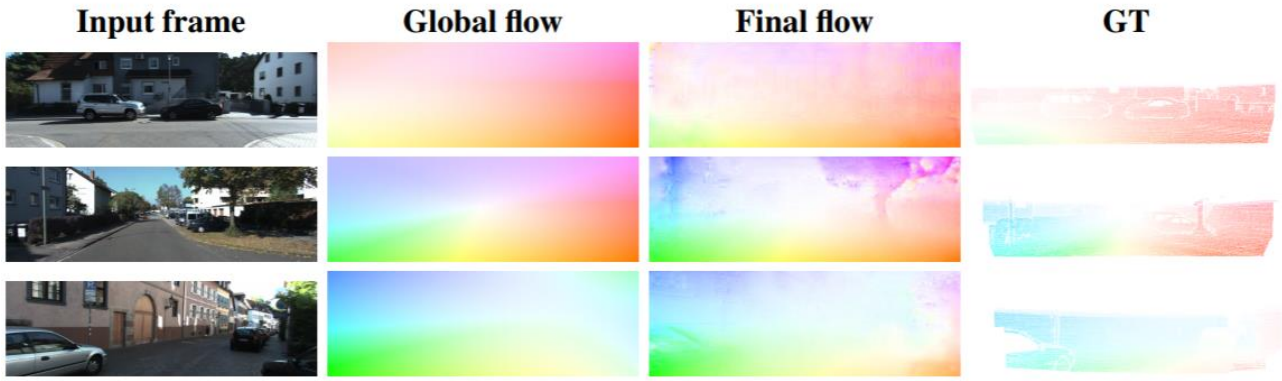
The KITTI Flow 2012 dataset was used for performance comparison, though the network was trained on a set of image pairs sampled from the KITTI raw dataset (dataset includes 44,000 frames acquired in the city of Karlsruhe). The table is divided into three sections: hand-crafted and supervised methods, unsupervised methods, and TransFlow with and without bilateral filter. After performance measures, execution times for each method are also reported. TransFlow compares favorably against hand-crafted and supervised methods, though not reaching the results of DeepFlow or EpicFlow.

"The most notable results of TransFlow are the 3x performance improvement on unseen data, which demonstrates that the method learns important transferable representations to compute the flow"

The authors and ourselves view the following as the most notable results of TransFlow: the 3x performance improvement on unseen data, which demonstrates that the method learns important transferable representations to compute the flow. It proves its great generalizes capabilities - where hand tuned methods had over fit to the dataset.

Method	Training		Testing		
	Acc@5	APE	Acc@5	APE	Time (s)
HoG	0.455	9.68	-	-	-
KLT	0.702	8.16	-	-	-
FlowNetS [9]	-	-	0.630	5.0	0.08
DeepFlow [22]	-	-	0.927	1.5	17.0
EpicFlow [17]	-	-	0.912	1.5	15.0
Long *et al.* [16]	0.716	4.70	-	-	486
Yu *et al.* [23]	-	4.30	0.652	4.60	0.03
TransFlow	0.857	3.335	0.692	3.90	0.14
TransFlow+Bilat	0.866	3.132	0.705	3.60	0.15

In the table above, Accuracy@5 is the ratio of motion vectors with end point error lower than 5 pixels; APE is the average point error of all motion vectors.



"Lack of generality mainly due to the synthetic nature of the rendered scenes"

Qualitative results on the Kitti 2012 dataset:

TransFlow was further experimentally evaluated on the Virtual Kitti dataset. Though featuring the typical automotive perspective, due to being synthetic it has unique problems, such as the presence of the typical artifacts of computer rendered scenes. Nonetheless, it is currently the biggest dataset providing ground truth automotive optical flow. Due to the dataset being very recent, no results are publicly available and we evaluate DeepFlow, EpicFlow and FlowNetv2. The results show how all three methods perform quite poorly on this novel dataset, demonstrating the lack of generality mainly due to the synthetic nature of the rendered scenes.

Method	Avg.	Night	Rain	Day
DeepFlow	48.10	36.53	52.93	54.84
FlowNet2	48.72	35.56	54.03	56.56
EpicFlow	49.77	36.65	55.60	57.05
TransFlow	4.38	2.71	4.91	5.52

More results, including the DR(eye)VE dataset with 555,000, featuring steep changes in image conditions due to transitions between day and night, weather conditions and scenarios, can be found in the paper.

Sum-up:

TransFlow is an unsupervised optical flow method that can be applied without requiring ground-truth generation. It successfully deals with the complexity of unsupervised training by first producing a global estimate of car motion (H-transform), using a shallow network. Then, using that product as initialization for a dense, complex network producing pixel-level transformation (F-transform).

The experimental evaluation demonstrate how the proposed method outperforms recent unsupervised methods while maintaining the advantages of simplicity and speed in an end-to-end framework build on neural network only.

"TransFlow is an unsupervised optical flow method that can be applied without requiring ground-truth generation. It successfully deals with the complexity of unsupervised training by first producing a global estimate of car motion using a shallow network"

"The teacher should adapt"

Sanja Fidler is assistant professor at the **University of Toronto** and a cofounder of the recently opened **Vector Institute**.

Sanja, what is the Vector Institute?

It's a research institute that we opened in March, and it's focused on fundamental research in the area of machine learning, specifically deep learning. It's government funded, and there are a lot of companies that contribute as well. We do pure research. Everyone can do whatever they want in different areas like computer vision, machine learning, NLP, and so on. We have many different topics. We hire around 20 faculties like research scientists that can create their own groups of students that they can hire. The idea is to foster what Toronto already has and bring it to the next level.. It is known for deep learning.

"Keep the talent in Canada"

What is the purpose of the group?

Originally, the idea was to establish research and keep the talent in Canada. Canada has less industry and less academic possibilities than the US, which is larger.

There is less focus on Canada.

Right - The idea was to keep all of these amazing, talented students in Canada.

That raises two questions. First, how did they get your talent, since you are originally not from Canada? The second question is how did it fall on your shoulders to found this institute?

Originally, I am from Slovenia. The first time I came to Canada was for a Postdoc position in 2011. My PhD work was very related to deep learning and theoretical representations of objects. I thought that would be a really nice place for me to do research. I didn't know much about Toronto, but I thought it would be good. I arrived in January, and it was so cold!…. But you dress appropriately, and you get used to it.

What was your drive to go on an adventure on the other side of the world?

Between my family and friends, no one ever left. Then on the last year of my PhD, I was invited by Professor Trevor Darrell to visit his lab. I went there for 7 months, and that was just amazing.

What convinced you to stay?

It was really the group. I really connected

"In Toronto, everyone is coming. There is always someone you can talk to who is interested in your research. It's awesome!"

with the group. It gave me the opportunity to talk to other researchers. The scale was much different, and people were much more engaged. These universities are structured in a way that allows people to be great. There are a lot of faculty and visitors. You are exposed to cutting edge research all of the time.

A research community like that can attract people to stay.

At the University of Ljubljana, the group was maybe a couple of students. We never really got visitors. You're kind of there on your own. Maybe you go to conferences, but that's it. In Toronto, everyone is coming. There is always someone you can talk to who is interested in your research. It's awesome! I said that's it. I'm going to finish my PhD, and then I am going abroad.

What was the most exciting part about moving?

I really love research. The opportunities there are just incredible. It's my passion. I can never switch off my brain. Even when on vacation, I am always thinking about my work.

Did you sacrifice anything by moving?

I really miss my family, but I find ways to see them often. I go home basically after every deadline. Before a deadline, I work really hard, and then the next week, I visit home and relax. I get to see my sister and her kids, my parents, and my friends.

This brings me to my next observation: if this impressive research community could attract you from across the world then maybe the Vector Institute can also bring in talent from all over.

Yes - That's our hope!

At what moment in your career, did you start to feel less like a student and more like a teacher?

Ooh - That's a tough one. I still feel like a student, just a different kind of student. When you are a student, you are learning the field. Now as a professor, I am learning how to teach. I always feel like I am learning.

"...deep teaching..."

What is more difficult, deep learning or "deep teaching"?

[*laughs*] Probably deep teaching... Deep learning is really interesting so it's easy to pick up.

What is more satisfying, a successful paper of yours or from a student?

[*replies with certainty*] A student's - The best thing is to see the paper being accepted and seeing the student become super excited. That's the moment

Apparently, coolness is a family trait: Sanja with niece Ajda and nephew Marsel

that makes it all worth it.

Will you have the same satisfaction in 20 or 30 years after teaching hundreds of students? How can you keep the spark of excitement in years to come?

I don't know about the future, but so far I feel the same excitement that I felt in the beginning of my PhD. When's there a new idea, I tremble with excitement. I enjoy collaborating with and teaching students. It doesn't seem like it's going to go away.

You seem very upbeat. What advice can you give to people whose papers were rejected or to those who are still waiting to have their papers accepted?

I feel that good research is always going to find a way to get published to become visible to people. If you really believe that you are doing something great, who cares about what a bunch of reviewers say? Maybe there is noise in the process. Maybe your paper actually isn't ready. Sometimes you need to agree with the reviews and make your paper better for next time based on their feedback. Next time it is going to be ever better! The key is to stay upbeat. You should not taking it personally. You need to believe in yourself. Don't get depressed from reading the reviews. Sometimes you get reviews which can get pretty nasty. This happens, but you need to remember why you still believe in your work. You can also learn something from the process.

Some people get stressed.

I guess for the first paper. For me, it's not stress, but it's more about feeling super curiosity about what will happen and feeling excited. You should not be stressed.

You have had a high percentage of papers accepted by the conferences. How does it feel to see you work succeed?

I always wait to submit papers until they are ready. Then you have a higher chance of it being accepted. Although then it can put pressure on yourself for the future. You can't always compete with your past achievements. If you did really well this year, you might want to do even better the next year. It can cause stress.

"I just really love what I do. My passion comes from a place of curiosity"

Are you more competitive with yourself or with others?

I am definitely more competitive with myself. I wouldn't call myself competitive. I just really love what I do. My passion comes from a place of curiosity.

After the CVPR oral (given by Lluis Castrejon on the left), which got the best paper honorable mention. On the right are Kaustav Kundu and Raquel Urtasun

It seems like you have many goals.

Yes - I set high standards and work to be as good as possible. I guess it might impose some stress on the students.

Did you ever see a talented student quit?

Yes, actually I have. That is the most frustrating part of the job, I'd say. There are two cases that I can think of now. One quit because of personal reasons, not because they were unsuccessful. His wife couldn't find a job, and they had visa issues in Canada

U of T scientists create software to analyze outfits

New program, which they hope to turn into an app, determines whether an outfit is stylish and offers suggestions.

University of Toronto researchers Raquel Urtasun, left, and Sanja Fidler are creating an app that assesses clothing and recommends how to be more fashionable. (STEVE RUSSELL / TORONTO STAR) | ORDER THIS PHOTO

at that time. There weren't a lot of job opportunities in Canada back then so they wanted to move to the US. Then he found an industrial job. He was a very good student so I'm still trying to get him back.

The other student was really, really talented. He said he wanted a taste of the industry before deciding to do the PhD. He might still come back. He went to work for a product team. It was a surprising choice because he has a lot of talent.

Students go through processes that can be quite frustrating. I think they want a taste of something different. In comparison, industry offers a more stable life than research.

What drives students to stick with academia?

Through these years, you learn a lot about yourself. You realize your limitations and boundaries before discovering what you really want to do.

What did you discover about yourself?

[*laughs*] Ahh, you're putting me on the spot! The learning process is never-ending. I didn't always know what to do. I didn't know if I wanted to work in industry or become a professor.

What convinced you to stay in academia?

In the beginning, I wasn't always sure. When I started, I was very afraid. I didn't know if I would be good at teaching and guiding students. I knew I wanted to try it. After the first year, I really enjoyed it, and now I cannot see myself doing anything else.

You have had extraordinary teachers. Which quality do you admire the most in your own teachers?

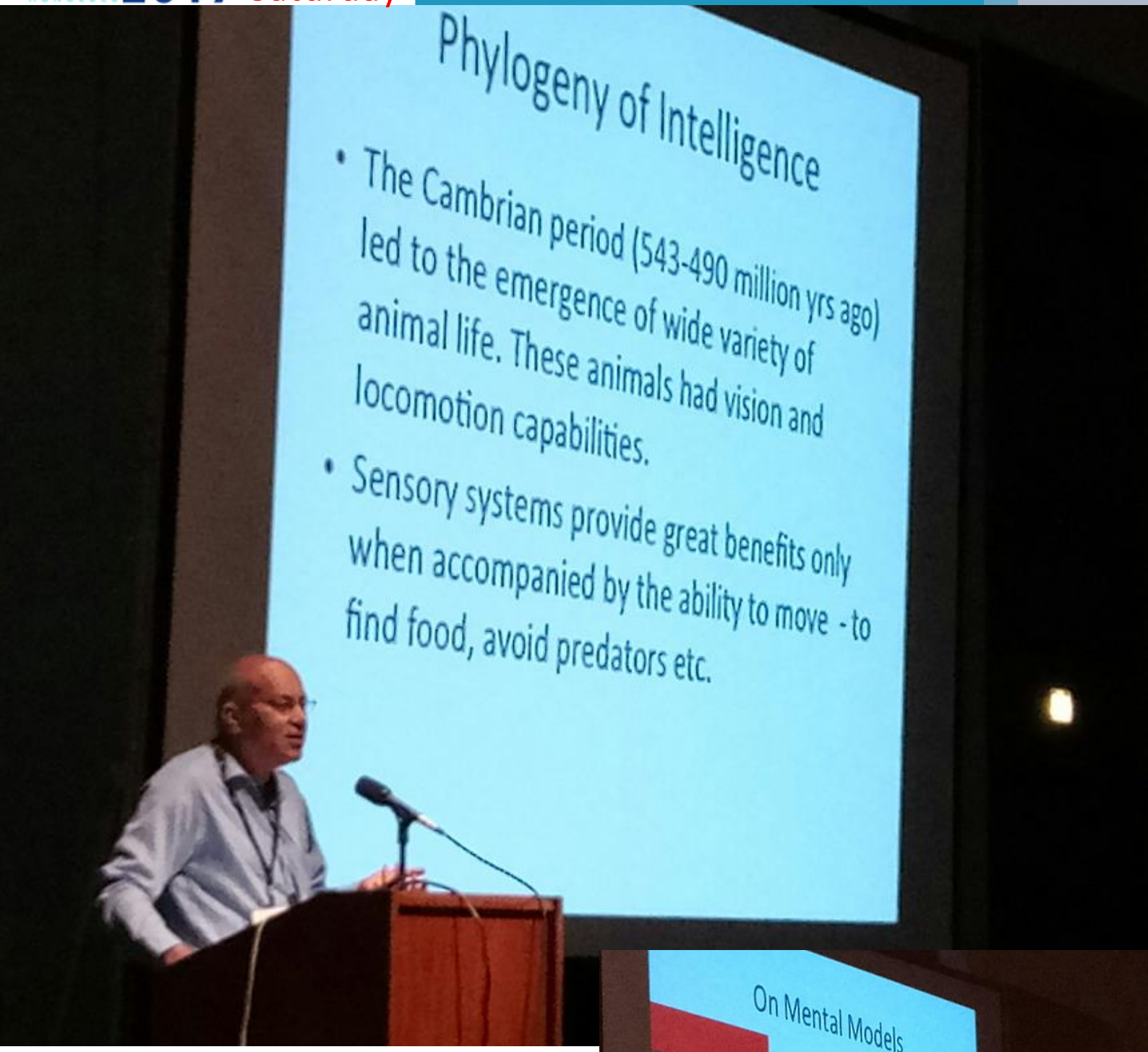
I had a lot of teachers that I learned from even from when I was a little kid. As a supervising student, I learned the most from **Raquel Urtasun**, who is also a Co-Founder of the Vector Institute. She guides students in a really natural way. She teaches them how to learn and how to approach problems. I am a little bit more chaotic. I tend to throw many ideas out there. Perhaps it confuses students. I learned that you shouldn't rush into things. You should go slowly and help them realize things by themselves rather than just by telling them. I think that is probably the best thing I learned from my teachers.

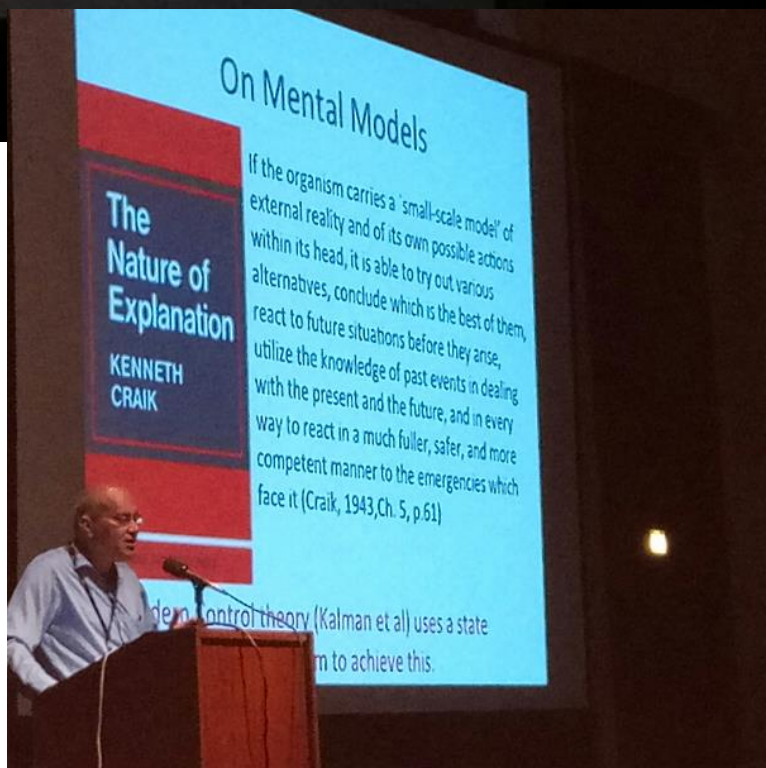
What is the most precious thing that you learned from your students?

[*laughs*] Wow, I've had so many! Every student is very different, and every student needs a different type of approach. Some like things to be very structured. I like to brainstorm so maybe I wasn't as structured. It changed the way that I interact. If I have ideas that I want to convey then I do it slowly.

Have you seen benefits of this?

Yes - I've seen progress, and a lot of projects are going well. Again, some people like it one way, and some like it another way. The teacher should adapt.





It is always passionating to follow a lecture given by Jitendra Malik. This year, at the Deep Learning for Robotic Vision workshop, he shared with us great quotes, good humor and new insights from his recent work with Google.

ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes

Angela Dai, a fourth-year PhD student at **Stanford University**, will give a spotlight talk today on **ScanNet**, a large scale RGB-D dataset of richly annotated 3D reconstructions of indoor scenes.

"It usually takes about five non-experts to get a good annotation per image"

Angela told us that "*the idea is that we want to power data-hungry machine learning algorithms like deep learning on 3D data*". In 3D there is more information than what you get in 2D because you have scale, and you know how far away everything is from each other. However, since 3D data is more difficult to capture than images, and because it is also more difficult to annotate, there doesn't exist much 3D data.

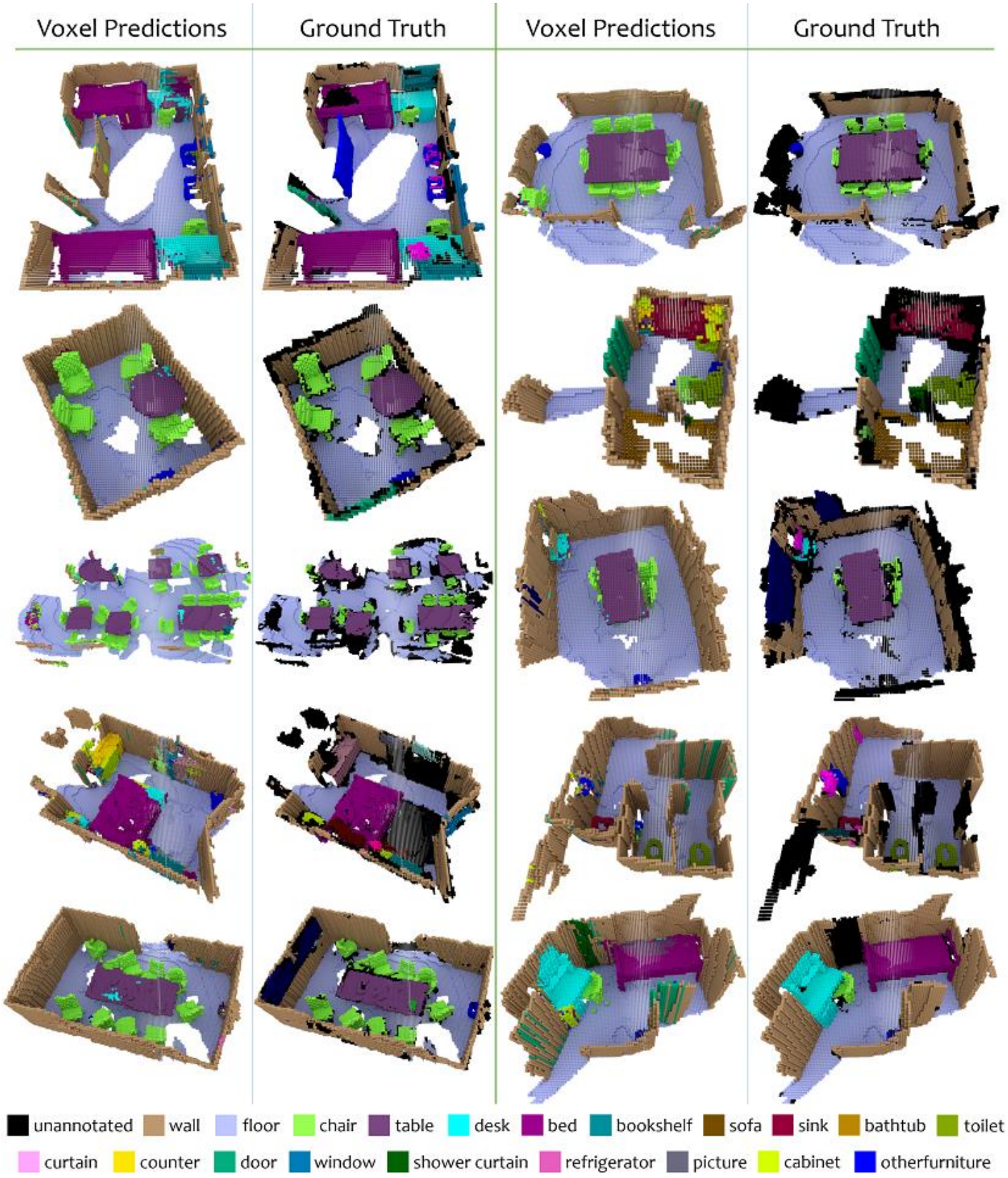
With ScanNet, they first wanted to build a scalable data-acquisition framework. This means first collecting the 3D reconstructions and to then annotate them in an efficient way, to be able to collect more than thousands of these scans. In the current version they have about 1,500 scans (video sequences of RGB-D), which they collected with users that were equipped with an iPad app and a depth sensor attached to it. After the videos are collected, they are uploaded to the servers, where they are automatically reconstructed. They are then pushed to an Amazon Mechanical Turk interface, which Angela and her team used to crowdsource the labelling of semantic segmentations. The task (which is usually done by non-experts) is that given a 3D mesh of a scene, to paint over this mesh in order to label instances. E.g., paint over a chair, a table, or a computer to tell what the objects are, and where they are in the space. Angela told us that it usually takes about five non-experts to get a good annotation per image. This can then be used for ground truth information when training for tasks like object classification: you cut out one of the objects that was labelled, and you try to train an algorithm in order to determine what object it is.

The idea of the ScanNet dataset is further to enable training algorithms directly on the 3D representation. For example, if you have a robot going around the room it should recognise what objects are in the room around it: you want to be able to identify not only that there's something three metres away from it, but also that this object is a chair.

Angela and her team also have several scene understanding benchmark tasks

on real world data. "*Because right now the only available large-scale amount of data for 3D is synthetic datasets*", Angela says, "*but that doesn't look anything like real-world 3D data*".

"Enable training algorithms directly on the 3D representation"



She says that usually if you train something on synthetic data, it is going to not work as well on real data because it hasn't seen data characteristics like the real world. "*The real world is noisy and partial*", Angela says, "*because you can't get a fully complete view of an object*". They therefore have some benchmarks to show that if you train for real data tasks, you can do much better than on training just on synthetic data. "*So it matters a lot that we have real-world data, and we could definitely still use more*". With ScanNet and the current 1,500 scans they provide a good start, and Angela told us that they were able to show that they are able to generalise to some previous real-world datasets that were smaller. "*But of course, we would like to get more, and this is still what we are working on*."

"A semantic understanding of the scene"

In the future, Angela would like to see something running on a tangible thing. She previously worked on 3D reconstruction, where they built a real-time 3D reconstruction system, but from this work she knows that it's very hard to use this kind of thing in practise. One of the things that is missing, she told us, is to be able to get a semantic understanding of the scene. Because even when a model looks reasonable, you still want to know where things are, or what they are, in order to actually at least have virtual agents or robots interact with them. "*I want to be able to make this happen for real scans!*", she adds enthusiastically.

Angela also told us about the next steps in line of this work. One of the "obvious things" is to scale this up even larger than thousands of scans - they aim to go up to ten thousands.

"What are objects?"

This however requires a different kind of data acquisition she noted, where instead of only crowdsourcing the annotation task, they also want to be able to crowdsource the reconstruction task as well. Besides this, there is also a lot to be done in terms of semantic segmentations. "*Right now, our tasks are still basically: What are objects?*", Angela explains, "*and there is a lot more interesting tasks on this type of data*".

One of them she is particularly interesting is connecting the real-world data with synthetic CAD models. They did this a little bit with ScanNet, but they want to push forward, to have an association with synthetic CAD models with the real-world scans. E.g., when you align a synthetic chair on top of the real chair, and then correlate these two. "*Ideally, you can basically learn a transform to go between real to synthetic*", Angela says. And this is a way you can make a model useable - since synthetic models are easy to manipulate and they are fully complete. It is also much easier to train something on synthetic data, but it's not easy to transfer that information to the real world. But if you had this correlation between the two, then it could be possible to learn the transfer between synthetic and real data. A method like this might be usable in a VR/AR application.

"Making your robots understand the world"

"*The big emphasis really is that we are interested in empowering semantic understanding tasks on 3D, which I think is a very important problem for actually making your robots understand the world*", Angela concludes.

Data and code are online here.

Angela presented ScanNet at a CVPR2017 spotlight.

Vittorio Ferrari

BEST OF CVPR

CVPR DAILY
July 21-26 HONOLULU 2017 Tuesday

Vittorio Ferrari is a professor at the **University of Edinburgh** and a research scientist at **Google Zurich**, in both of which he runs a research group.

"I salute the students that are braving the new world in these days."

Vittorio, do we have in common an Italian background?

Almost right - I am Swiss, from the Italian part of Switzerland.

So we are neighbours! I am from Milan, and there we say that we are nearly Swiss.

Actually, in Lugano, I have been told by the Swiss Germans that I am nearly Italian. [*We both have to laugh*]

So you grew up in the Italian part of Switzerland, then decided to become a scientist and your career brought you to different places.

Yes. Interestingly enough, when I was a kid, there was no university in Lugano, my hometown. So even just to study, before being a scientist, you had to move. That started my journey about half a life ago.

Where has this journey led you, what are you working on now?

I work on various problems in computer vision, but my general life mission since 5-6 years is to try to learn computer vision models that are able to localise objects in images with high quality and least human intervention possible at the same time. This sometimes is described as weakly supervised learning. Recently I have been working a lot on human-in-the-loop learning, where there is a bit of human supervision and intervention during training. For example, as a

typical outcome you would get a model that is capable of labelling every pixel in the image that is containing an object, but at training time you perhaps only need image-level labels. And yet, you can get a localisation model almost out of thin air.

Why are you so passionate about these kind of problems?

There are two really good reasons. One is an intrinsic, scientific reason and the

"Oh, I am more passionate now!"

other is the impact it can have. The impact it can have is that we will one day be able to train from tens of millions - no, billions - of training examples that cover tens of thousands of object detectors. This is in fact necessary to reach human-level ability, you need lots of samples and lots of classes. And one day we will be able to do that at a cost that is within the ability… maybe not of everyone, but at least within the ability of a millionaire [*he laughs*], in the million dollar range. Now this would be absolutely impossible, if you want a complete annotation of every pixel in an image. You cannot do a million objects, basically. The problem will only be solvable once we have all the annotated data, and we will not annotate it by hand the way we are doing it in the fully supervised world. That's why reducing the annotation time is not just a sport, it's an enabler of solving computer vision. Now if you go to the scientific reason, which I am even more passionate about, it is a very interesting information-theory type of trap. When you have a weakly supervised learning problem, let's say, this image where we stand now: this is a couch, this is Ralph, this is Vitto, this is a plant in Hawaii. You have these labels, and there is actually combinatorially many assignments of the pixels in the image to these labels. And all of them are consistent with the labelling of the image, but some of them make more sense in terms of regularity. It's very interesting that theoretically, there are many solutions that are valid, so that strictly and information-theoretically speaking, it is impossible to reconstruct pixel-level labelling of an image from image-level labels. And yet, there exist some assignments that are more likely to make sense in the visual world. For instance, all the pixels on your face probably all take the same label, they're all face. For me it is very exciting that although we know that there is no perfect closed-form solution that will work, there is certain families that make more sense in the visual world and that lead to good results at test time. So somehow I like the fact that you start by saying that the problem is impossible, and yet you try to solve it.

"Like a kid in a candy store…"

You sound still as passionate as when you started to study…

Oh, I am more passionate now! When I started my PhD, I felt like a kid in a candy store. You jump at everything that looks cool, and you grab something, lick it a bit, then you take something else… so there is no continuity of mission. Now I am equally motivated, but because I focused the energy of my team over multiple years on a family of problems, I also see a lot more progress. And I appreciate the fine details of these families of problems. So in fact I actually feel more passionate now compared to when I started.

Do you have tips on how to keep the passion over a long period of time?

I started my PhD 17 years ago, and one way to keep the passion over 17 years is to change the family of problems every once in a while, to freshen up. Stay in machine learning and computer vision problems, but change at the beginning every 3 years and later every 6 years, and eat from the diversity of the computer vision fruit - it's a really big fruit.

Is that something that others do as well?

If I may dare to mention the really great, like Zisserman and Malik, who manage to keep the passion for a long time, they all have one distinguishing mark: they worked on a lot of problems, and typically they make one landmark contribution in each era.

Can it happen that somebody enters into a field and realises that they shouldn't have?

Oh, absolutely. It happens when you're younger, and it happens when you're older. You have to be able to feel whether putting energy into an area is going to lead to things you want. Which are always the same two: happiness for yourself, so you have fun, and the second is your publishing and that people are interested in what you write. These two criteria are often in contradiction. So you need to feel it as fast as you can. I would say if you're not happy after six months after entering a field, you should change.

How do you rebound in the case it doesn't work?

When you are the first implementer - a PhD or a postdoc, before you are a professor, rebound is somewhat easier. You have to have the self-discipline of going to your advisor and saying, look, this thing doesn't work. And then normally it's about having a picture of the new thing. So just saying "*I hate what I'm doing, and I quit because it doesn't work*", then the alternative is the void, and the void scares everybody. So you need to set aside some time. Normally when I felt I wasn't doing so well in an area, especially when I was younger, I would just say: this week, I only read. I don't program anything. And just read as diverse stuff as possible, and then decide what to work on. When you are a older and you are a group leader, then it's harder to rebound, because you are very much in love with your own vision [*he laughs*] and you don't quite see why it doesn't work.

And you have a responsibility for the people who are with you.

Yes, telling your students: you know what, because of various reasons, perhaps technical impossibility reasons or because somebody else already implemented your idea, you have to change direction. This is tricky, but it's important to do it. As you said, you have a responsibility for the student. And sometimes the best interest of the student is to radically change topic. As a group leader you must make these choices.

"humble down sometimes"

Might it also be the case where the student opens the eyes of the professor and says that something is not going to work?

Absolutely, sometimes it's exactly the PhD student or postdoc that has to go to his boss and say: you know, Vitto, this thing you like so much - it ain't gonna fly [*he laughs*]. The professor

"The noise is noise with respect to the center point, but it's signal with respect to scale"

has to be able to humble down sometimes. Sometimes the first implementor, the person that is doing the actual work, is actually not seeing that it could be working, and the group leader sees that it could be working. And then the group leader should stay on track. But sometimes the group leader is just illusioned, is lost in his or her own ego, is in love with the own idea, and then you have to say: you know what, you're right. And this dynamic between bottom-up and top-down leadership, it is a big feature of a healthy group.

Do you think there are any changes between your generation of students and the generation of students that you see now?

Ok, ehm… Temporally skip. I give you just a quick comment, we can continue after. Just off the record comment… this is such a awesome series of questions! [*we both laugh*] It's just so fun. Ok.

[*Laughing*] **Why don't you want to put this into the interview?**

[*Still laughing*] Ok, put it in. This is so cool. So before we re-start… what was the question again?

[*I repeat the question*]

Ok, this is an awesome question! See, it's very interesting. Let me first say how the environment changed. When I was a student, everything was much slower. You had an idea, and you thought: it's awesome! Then you had approximately a year, or a year and a half, time from an idea to a publication because the density of people working in your area was low and, you know, in the end you wouldn't be scooped. These days, if you work on something hot, especially on neural networks understanding something - the very middle of the field - between an idea and somebody scooping you, you have maybe six months. So it's becoming more stressing. But at the same time, because it's so much denser, if you do something good you get a lot more citations. And citations are a big currency, a big mark of success, that you trade for positions and professorships. So in a way it's harder and easier, at the same time. But it's certainly more stressful, and I salute

the students that are braving the new world in these days.

Now, in terms of skills. I believe that this generation is able to get somewhere faster. Nowadays, they have more tools, and they can recombine software pieces. So in a way it's exciting, the pace at which they can go. Perhaps if I can dare to make a recommendation, something that I felt that back in the days were perhaps a little bit better. The students today have a tendency to be very rushed to say: I am working on whatever is hot now, and what happened three years ago is forgotten. And this is very short-term sometimes. So perhaps back in the days, the students were trying to think a little bit more like: how can I change things globally? And they looked a little bit more beyond their field. So perhaps this has changed, but this is also a reaction to the environment. Today things just have to go quicker.

"How can they be so silly, or what a genius..."

What is the biggest surprise that you ever experienced from a student?

Oh, voilà! I would need a lot of time to answer, because there were so many times my students surprised me. So many times! Sometimes positive, and sometimes negative. But they often surprise you. And it's very important - back to information theory - anytime a student surprises you, positively or negatively, and you think - how can they be so silly, or what a genius - both times, take a step back as an advisor and update your own neural network in your head [*laughs*]. Update the student model, because that's where you learn, the surprise. So, I will just answer with something that comes to my mind which is fun, it's part of our papers we have at CVPR.

"That's awesome! How did you think of this?"

It's a technical contribution, but I thought it was really fun. My student was working on this project, where we try to learn object class detectors by annotating objects using the center point, instead of drawing a box around it. And my student was saying: you know, we should ask two people to click in the middle. And I said, forget it! It's useless! It's just a little bit of noise cancellation. The student said: well, you know what, Vitto. If they are both asked to click in the middle, they are going to make an error. And I said, so what? So the student said: but the errors they make is related to how big the object is. Because if the object is big, the two annotators are going to click further apart from each other, and on the smaller object they will click closer. And therefore we can estimate how big the object is based on the errors the annotators make. And I was like: that's awesome! How did you think of this? We are exploiting errors to get information about the object scale out. And in weakly supervised learning, object scale is one of the big holy grails. If you have it, it makes it a lot easier to learn. And so, you know, we dumped it into this paper and it became one of the coolest bits of the paper. I thought - how did you think of exploiting the errors in humans? When I think about it, I want to cancel out the errors, not turn them into information. I was really impressed by the student when he said that. It was very clever: the noise is noise with respect to the center point, but it's signal with respect to scale.

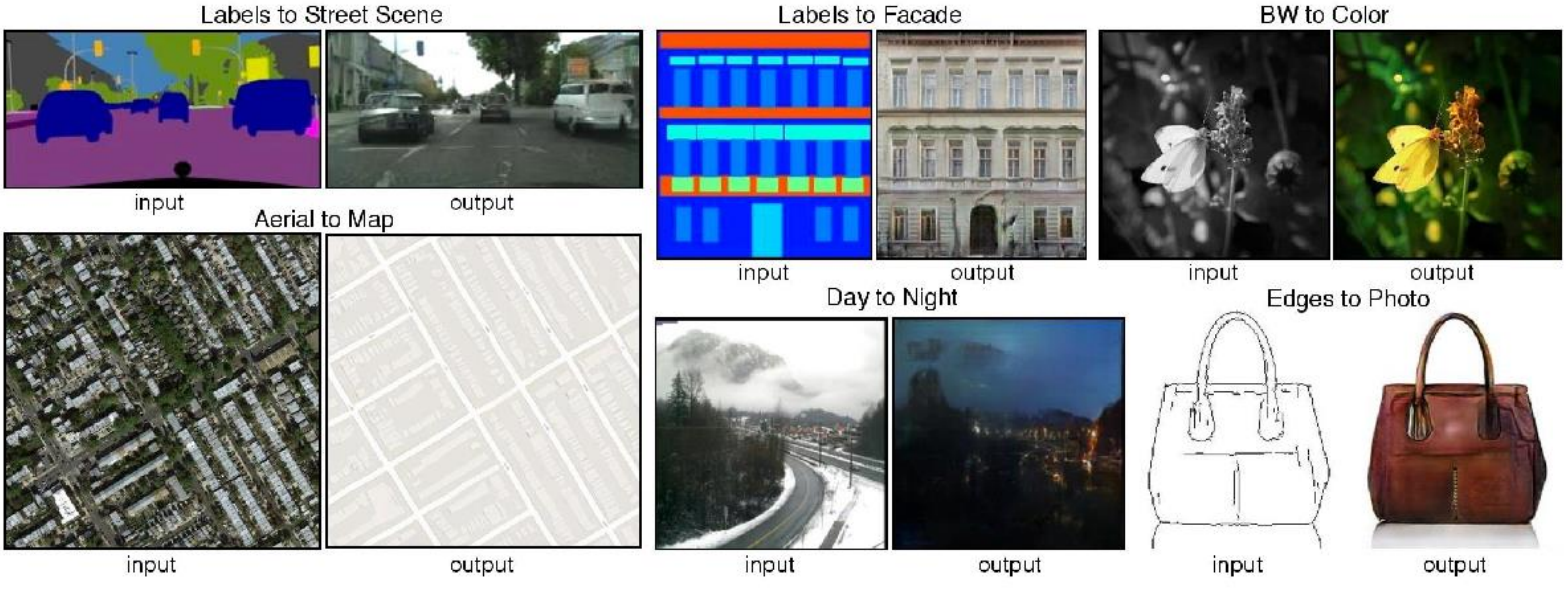
Image-to-Image Translation with Conditional Adversarial Networks

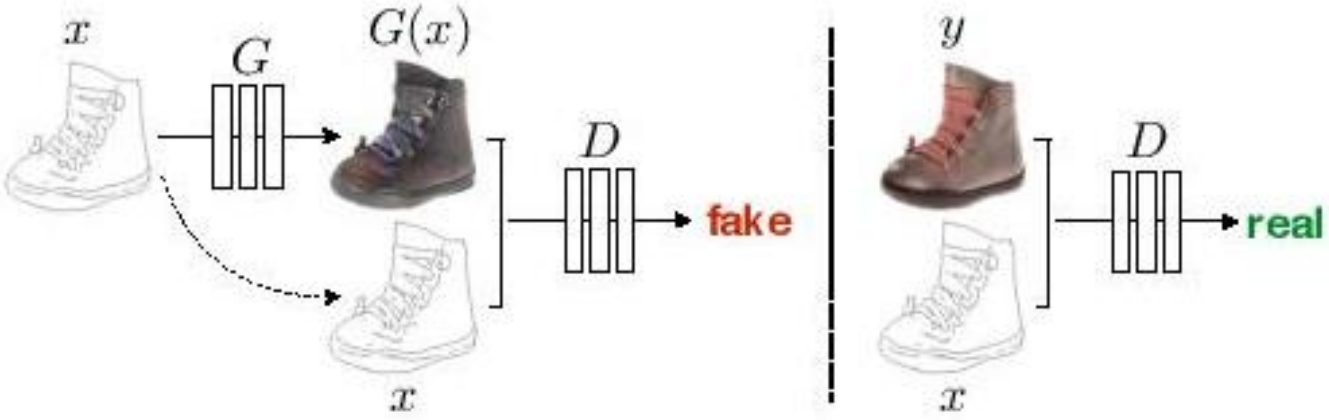
Phillip Isola is a postdoc with **Alyosha Efros** at **UC Berkeley**.

"There's a lot more problems that are conditional than unconditional, especially practical problems in computer vision and graphics"

Phillip Isola presented his paper "**Image-to-Image Translation with Conditional Adversarial Networks**", which is joint work together with Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Their idea is to use **generative adversarial networks (GANs)** to solve image-to-image mapping problems, and in their paper they demonstrate that these are a general-purpose tool that can be applied to a lot of problems.

GANs, which were introduced by Ian Goodfellow et al. in 2014, and are a popular idea at the moment, and a large part of our community has gotten quite excited about them - "*rightfully so*", Phillip says. He told us that previously a lot of people have done work on unconditional GANs, which were used to generate random images. But Phillip and his co-authors thought that it might be more compelling to look at the conditional case, where you use a GAN for regression problems to learn a mapping from inputs X to outputs Y.





"*There's a lot more problems that are conditional than unconditional, especially practical problems in computer vision and graphics*", Phillip told us. For example semantic segmentation or edge detection are both conditional image-to-image mapping problems, or things like image colourisation (taking a black-and-white photo and producing a coloured version of it). In all of these problems you want to learn a mapping from pixel-to-pixels, i.e., images-to-images.

Phillip explained to us that what happened in the last couple of years is that CNNs have turned out to be a very generic way of processing images and are used for a lot of problems. But usually a CNN is only modelling structure in the input space. CNNs with the standard regression loss are treating every output pixel (of the semantic segmentation map or edge map) as conditionally independent given the input, so they don't model semantic structure in the output space. As a reaction, the community has already done a lot of structured regression problems modelling structure in the output space, for example using conditional random fields. But what Phillip's current work is doing is using adversarial discriminators as a way of learning a structured loss function to model structure in the output space. You thus have a neural network that models structure in the input space, and a neural network that models structure in the output space, to do generic things that can process images. "*A year ago this was all very new and unexpected. The field has developed these ideas all together and we are one of them.*" In their paper, Phillip and his co-authors show that this kind of approach is suitable for many image-to-image mappings, and they demonstrate that this works well on a lot of problems without any change in the architecture or method.

"We then realised that we could remove almost all the bells and whistles"

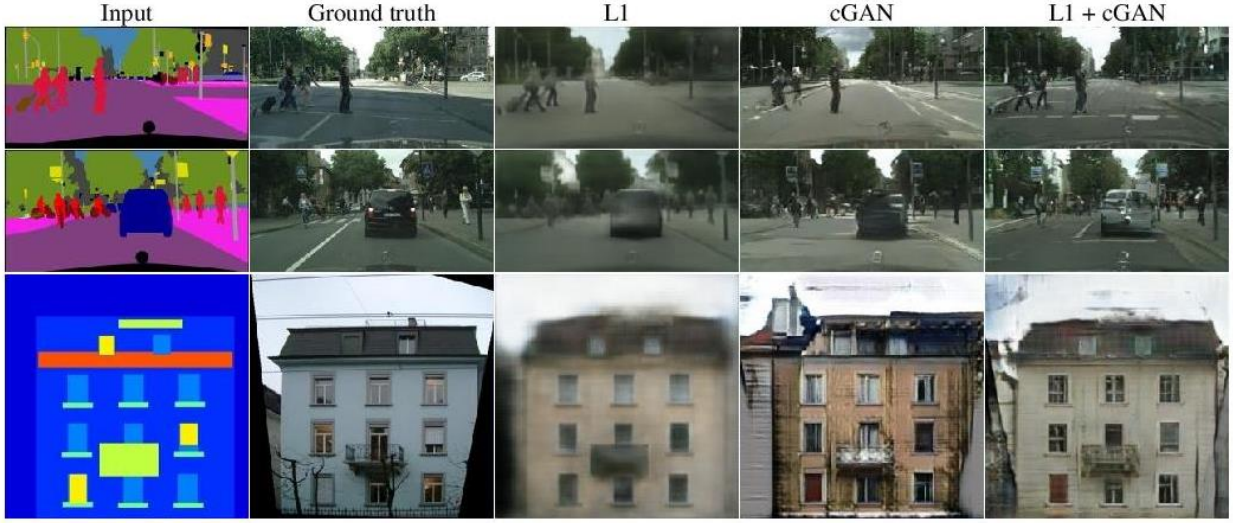
Phillip also told us about some insights they got from working on this problem: "*The process was that we added a bunch of bells and whistles and got something working, and then realised we could remove almost all the bells and whistles*".

The final model therefore only has a couple of tricks which turned out to be necessary. One thing they found is that while GANs are hard to optimise and can be unstable in the unconditional case, the conditional case is a lot more constrained: the conditional color distribution given a black-and-white image has much lower entropy than just the distribution over all possible random images. Because you have paired inputs-outputs, you are now also in a supervised learning setting: you can mix your GAN objective with the more traditional supervised objective. That's what they did in this work - they added L1 regression as an extra term in the objective to stabilise things. This leads to faster convergence and learning is more stable. "*The nice thing is that if you average in this L1 regression with a small weight, then it doesn't really change the final results*", Phillip told us, "*and you still get nice clean GAN-quality results*".

For future work, Phillip thinks there is still a lot of exciting things to do in the conditional GAN setting for image-to-image problems, and they already have a follow-up paper, called CycleGAN. Here they start with the observation that in the conditional GAN setting they needed paired supervised data. Given coloured images for example, they can train a mapping from black-and-white to coloured images in a supervised fashion, because there a million images to use for this. But if you want to learn a mapping between two domains, like paintings and photos, then you don't know the pairing. You might for example want to learn the mapping from a photo to a Monet-style image - but since these don't exist, this can't be trained in a supervised fashion. So without the paired data, you can't apply things quite the same way. "*But it turns out that some small changes allow you to also learn the mapping in the case where you don't have paired data, but you just have two stylistically different domains*", Phillip concludes.

Current paper *CycleGAN*

"The nice thing is that if you average in this L1 regression with a small weight, then it doesn't really change the final results"



What Is the Space of Attenuation Coefficients in Underwater Computer Vision?

Derya Akkaynak is a postdoctoral researcher at the **University of Haifa**, jointly with the **Interuniversity Institute for Marine Sciences in Eilat**.

"The other deep in computer vision"

Derya (whose name means '*ocean*' in both Persian and Turkish) told us about her work which she is presenting today, titled "**What Is the Space of Attenuation Coefficients in Underwater Computer Vision?**".

The poster she is presenting is about improving color reconstruction and color acquisition in images that are collected underwater with underwater robots or divers. Derya, who is an oceanographer and mechanical engineer (not a computer scientist by training), is working on understanding how light propagates underwater and how it gets captured on camera sensors, to find out how we can compensate for the colors that are lost, in an accurate and objective way. Her paper leverages decades worth of data from optical oceanography to improve underwater computer vision algorithms, bridging the two fields.

A main challenge of her work is to validate the mathematical models they build, and see if they actually work in an underwater setting. This requires many dives, a lot of equipment, and a lot of hardware and sometimes things can go wrong, or the results are unexpected, and then they have to go back to the model and adjust it. "*So it's a constant iteration between work on the computer, and work in the sea*", Derya says, which is different to most Computer Vision fields, where work is mostly done on a computer.

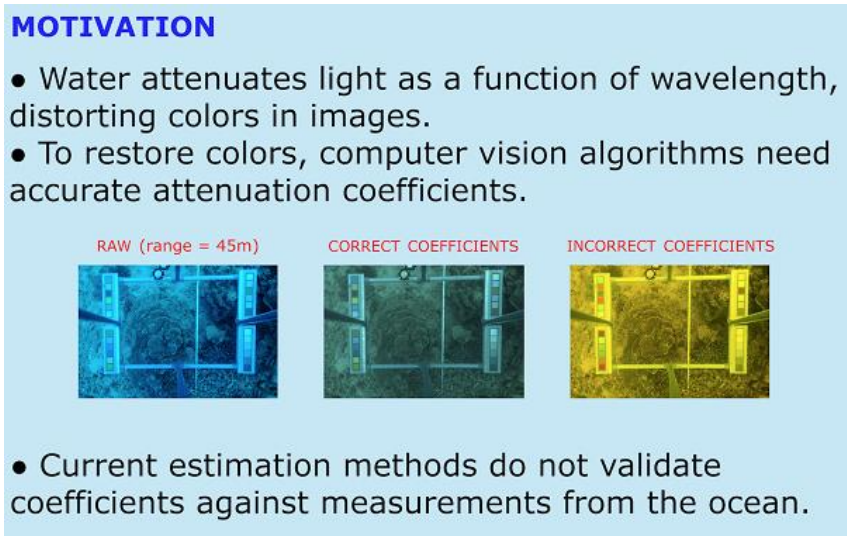
Talking about previous work, Derya explains to us that there is an existing system of equations for underwater image formation, and that everybody uses these equations. However, Derya and her co-authors looked at how these equations were derived and they found that due to two simplifying assumptions there are errors that are introduced that affect people's work in color reconstruction. So instead of using what was commonly accepted, Derya and her co-authors questioned it and were then able to highlight the weaknesses, and offer a better solution.

We asked Derya if she thinks that we can one day see underwater images just like we see things in normal life. She hypothesizes that this might be possible with a lot of specialised equipment, because we now understand very well what happens to light and how cameras capture light underwater.

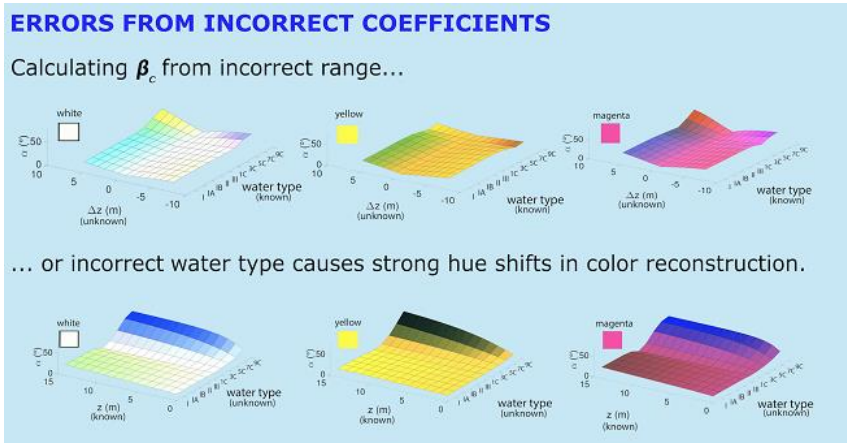
So it's possible to un-do that effect, but she is unsure if it will be commonplace enough to just put on a mask and that this mask will compensate for everything. She thinks that such a mask will probably complicate diving a bit, and might be more interesting for commercial applications.

"*Our next step is to now derive a new system of equations for underwater image formation*", Derya says, "*to compensate for the two weaknesses that we found*". They want to be able to tell people what kind of errors they should expect when they use the old equations, and they can then judge how good their results are when all the errors are taken out.

For their work, Derya and her co-authors used image formation equations that are known to be used for camera and image simulation.



"Our work has implications for those working with other scattering media, such as milk and wine"



What was important and key in their work, and their biggest contribution, is that they have brought over eight decades of knowledge from optical oceanology into Computer Vision. Because in Computer Vision, researchers estimated some coefficients that they use to correct colors. "*But they never checked if those coefficients actually make sense given ocean conditions*", Derya says. And in Oceanography it is known that light attenuation changes by place and by time, which oceanographers have mapped for the last eight decades with various instruments, from very simple to very technical. So what they have done in this work together with Tali Treibitz is to bridge Computer Vision and Oceanography.

Derya also told us about what she thinks is the biggest misconception about underwater imaging: "*In almost 99% of underwater Computer Vision papers I read, people state that light attenuated unevenly underwater; red attenuates faster than blue or green*". And this is exactly what happens in waters that are dominated by plankton (small plants that drift in the water), she explains. But if you go to coastal water, where there is contamination from rivers, soil, sediment and other non-organic substances, actually blue attenuates faster than red.

Derya presented her work at a CVPR2017 poster session.

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.

The Impending STEM Workforce Shortage: How to attract top talent, regardless!

Moderator: Ralph Anzarouth (Marketing Manager, RSIP Vision)

Panelists:

Dr. Terrence Chen (Sr. Director, Siemens Healthineers),

Dr. Tianli Yu (CEO, Morpx),

Julie Kemp (Principal Recruiter, A9.com, an Amazon company),

Dr. Xin Chen (Sr. Engineering Manager, HERE Technologies; Adjunct Professor, Northwestern University and Illinois Institute of Technology).

Ralph: Good afternoon everybody and welcome to our panel on STEM recruitment. The subject of this panel is recruitment in scientific professions. There are plenty of ideas about it, and I am very glad that we have a very diverse panel with persons which will help us to have a very enriching and passionate debate. I thank the public for being here and I thank the panelists as well. I would like everyone to introduce himself or herself so we can get started.

Tianli: I am Tianli Yu, the CEO of Morpx. We are a start-up company based in Hangzhou, China. We make toy robots using Computer Vision and AI.

Terrence: My name is Terrence Chen. I am a senior director at Siemens Healthineers, and I am located in Princeton, New Jersey. We are an organisation within the Siemens Healthcare Technology Center.

Julie: I am Julie Kemp, I am a principal recruiter at A9.com, which is a subsidiary of Amazon. I have been in technical recruiting for over 20 years and I have been focusing on computer vision hiring for the last 7 years.

Xin: My name is Xin Chen. I work at HERE technologies and I am based in Chicago. I am a senior manager, and we work on high-definition maps for self-driving cars.

Ralph: We often hear catastrophic predictions about the availability of technical talent in the next few years. Some people say we will lack millions of engineers and computer scientist. Do you agree?

Julie: Specifically in the US they state that there is going to be about 1.4 million new CS jobs coming up in the next few years, and the US universities can only handle about 29% of that. So we really need to think about hiring as a global decision. We have to focus on looking for where the talent might be, and hiring there and really seeking out the best people across the world.

Terrence: What I see from here from CVPR, the conference becomes much larger every year. At least from the domain of AI and computer vision I think there are much more students interested in these domains, and joining these domains. So in general for STEM I think there might be a lack of students or workforce, but for AI and computer vision I am very optimistic.

Ralph: I would like to follow up with something that Julie just said, on the use from the US of talent that comes from elsewhere. There are many overseas markets which find it quite

unfair that America is so good at attracting talent. What is your take on that?

Julie: The United States is a wonderful place to live, and it's a great place to come and work and to expand your capabilities. But for instance Amazon is really a global company and we do try to build our technology centers around the world so that we can attract people who don't want to transition to another country and give them the opportunity to work on really interesting problems.

Xin: Just within in the US, different locations are very different. So people, because of their personal preference or family, they want to be at different locations. At HERE technologies, in the US, we have offices in Chicago, in Boulder, and San Francisco. So this accommodates for talent and their interests. So for example Chicago is a very sophisticated city and a good place to live with families. In Boulder there is a lot of outdoor activities with a nice scenery. San Francisco is a very vibrant area. So I think this way we have offices with a variety of coefficients.

Julie: I think that also speaks to the fact that we recognise that it's hard to hire great people and we have to be flexible about where seek people out, and where they work. Because it's less about the employer demanding where you are going to be, as it is about where you want to work.

Terrence: Nowadays most of the companies are global, like Siemens as well. Siemens is originally a German company, but we also have a global presence and the size in the US is very big. We do a lot of things to encourage STEM force from school. We provide all kinds of tools for teachers, parents and also for students to engage them in the study of STEM topics. We also have a competition in technology and science, a fellowship and scholarships, to encourage STEM workforce in school. So all this is not just to take the graduate like a consumer, but really try to engage the development of the STEM workforce.

Competition

Ralph: The next questions connects to what has been said about recruitment possibility of big companies. Some of the smaller companies complain that if the big guys - and Siemens and Amazon are certainly one of them - are taking all the talent with conditions that we are not able to meet, they are not able to compete. I would love to hear what you have to say!

Tianli: As a start-up company we actually do have some disadvantage when competing with bigger companies in recruitment. I think what we use as an advantage as a small company is that we can let the people choose to work on more interesting problems. Our company is making AI toy and robots. A lot of people are attracted by the work

From left: Ralph, Tianli, Terrence, Julie, Xin

that is done. From a certain point, money does not have an effect anymore. It's more about how interesting the problem is you are working on. So I guess everyone has its own way to attract talent.

Terrence: With the current high competition with recruiting talents, even for a big company, it is not easy. In our case, we will need to compete with a lot of competition in terms of compensation and we need to really encourage our candidates with all the means that we have. For example, we work on healthcare topics. And we are trying to find people who can imagine themselves in revolutionising healthcare, with the technology we have. We are in Princeton, which is a top school district in the United States, and the housing prices are reasonable compared to some other areas. So if you are with a family and want a decent house, this is a good place. Things like that - you have to use all the weapons to get top talents. Sometimes you get them, and after a few years you lose them again. So I think this will never end.

Julie: I had the advantage to work both in a start-up and a large company, and I appreciated the experience I have had in both of them. I think the underlying thing that appeals to people when they are looking at an opportunity is: what's the work? If it's a really interesting idea or problem, that's where they are really going to gravitate to. Whether it's a big or small start-up, or a large company.

Xin: Our company is kind of in-between a start-up and a big company. So I think our advantage is that we have the right resources like big companies have, but also we have a start-up environment. So this is something appealing to many applicants.

Gender imbalance

Ralph: I want to change to a maybe more complex topic. I was told several times the joke that "in Computer Vision, we have more Davids than women". So the next question would be - what are we doing to have a workforce to be more equal?

Julie: This is a really hard problem. It's something that starts way earlier than when somebody is looking for a job. In middle school, 74% of girls are interested in engineering opportunities. But when it comes to graduating high school, only 0.4% of graduating women move into a Computer Science degree. So there is a huge break there and we have to understand what that break is and how we can improve on that. It's a societal issue in a lot of ways. It is very challenging as a woman to move into a science degree, but I think it's really important to have that diversity when it comes to innovation and research.

Xin: Just an example, we have female engineers and some of them have become leaders of the company, which I think is very encouraging for female colleagues. In our team, we have female colleagues in deep learning and they are just as good as the male colleagues. I personally have two daughters and hope that the landscape will change.

Ralph: This brings me to a very important point. How do we solve these problems actually? From my interviews I know that one answer comes again and again: the need for role models. It helps to show that even when there are difficulties, if

that person could overcome this, then I can do it too. But what about positive discrimination? Some women say they don't want to be positively discriminated - "don't give me special treatment". Others respond that in the ideal world it would be like this, but we don't have an ideal world so we need the positive discrimination. Who is right?

Julie: I don't want such a treatment as a woman. I want to be valued as who I am, and my ability to think through problems, my ability to solve difficult challenges. And I think any woman really is the same. I can't speak for every person in the world, but I do think that we love the fact that we can be valued as equals as much as possible. But we also have to acknowledge that that's not exactly the world that we are living in. I don't really look to specifically hire any particular woman, but I want to give the women we are talking to the opportunity to envision themselves here, at my company. And so we will make sure to add a woman to the interview panel, for instance. So that they can see that women can be successful here. So as you said, that they see mentors, that they see people similar to themselves and there is a path for them. But again, it's a difficult problem.

Ralph: Let's get back to the geographic and ethnicity aspects. For example, Africa is a huge continent - but how many people from Africa are here? How can we make sure that all the talent which could be available in the scattered areas in the world is included?

Julie: There a myriad of macro-economic and political issues that play into some of those regions that are really hard to get a foothold in. And those are things that are really difficult to solve and will take a lot of time. **But as you mentioned earlier: we need mentors**. Having those mentors that are able to mentor in specific regions, wherever they are, I think that's an area where this would be very helpful.

Ralph: We agree that we have a huge opportunity to reach more people in areas which we don't really consider today. This panel is coming to an end, and I would like to give you the opportunity for a closing remark.

Tianli: I want to say that I want more diversity in engineering because only with diversity, you can design products for everybody.

Terrence: There are two things I want to mention. One is through the preparation of this panel discussion I realised that how Siemens is doing for the workforce of STEM, and there is this website siemensstemday.com, you really need to go there to take a look. The other thing is in our company we have a lot of open positions, so we are looking for top talents, and you are welcome to join us.

Julie: I'd say in the largest scheme of things we really should think globally, and act locally. So think of this as a global problem that is larger than anyone of us can individually solve. But if we can make strides individually to helping improve the STEM education and helping improve the interest areas in girls, for instance, or underrepresented minorities.

Xin: Especially for this conference I think that the relationship between industry and universities are really important for recruiting and for training and to build our future talent. So I would encourage companies to have close relationships with professors to do research collaborations or supervise students. That will help a lot.

Women in Computer Vision

Nour Karessli is a computer vision engineer at **EyeEm**, who are located in Berlin. She published "**Gaze Embeddings for Zero-Shot Image Classification**" here at CVPR, together with Zeynep Akata, **Bernt Schiele** and **Andreas Bulling**.

Nour, where are you working at the moment?

I work at EyeEm, which is a photography company, and we work on cutting-edge technology for computer vision. We connect a community of talented photographers with iconic brands and sell photos. I finished my master's degree in July last year and started at EyeEm in August, so it's been a year.

What was the focus of your master?

My master thesis was about gaze embeddings for zero-shot learning for classification. I did it at the Max Planck Institute in Saarland.

I understand you did not start your studies there?

I started my master's there, and before that I was doing a bachelor in Syria, at the Damascus University.

"We make use of the human ability to distinguish between different classes unconsciously"

You are doing a presentation today. What is the work that you are presenting?

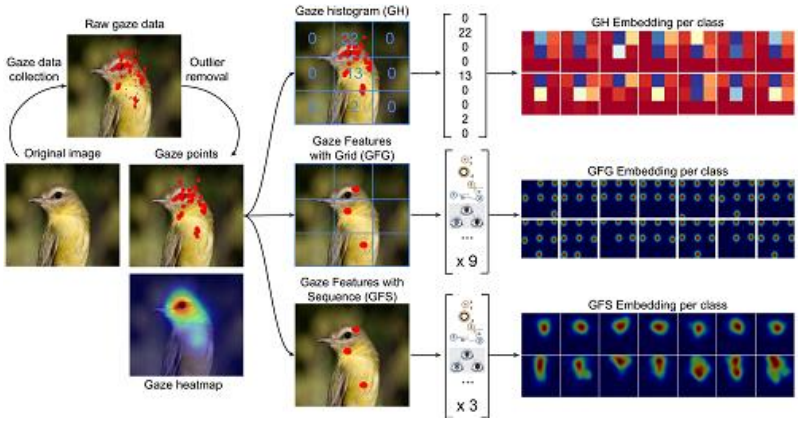
The work I am presenting is a paper about gaze embedding for zero-shot image classification. In this paper we use human gaze information to guide the classification task in a zero-shot setting. We make use of the human ability to distinguish between different classes unconsciously.

What is the novelty of this work?

Previous approaches used object discriminative properties collected by experts, and then the annotators had to go through the objects and annotate these attributes. This is very costly, especially for fine-grained classification. It's also difficult because the categories are visually very similar, and thus our suggestion is to use the gaze information. It's cheaper and faster, because it's implicit. You just ask the annotator to look at the image and distinguish between the objects, and then the human - without thinking about it - will focus on the important features.

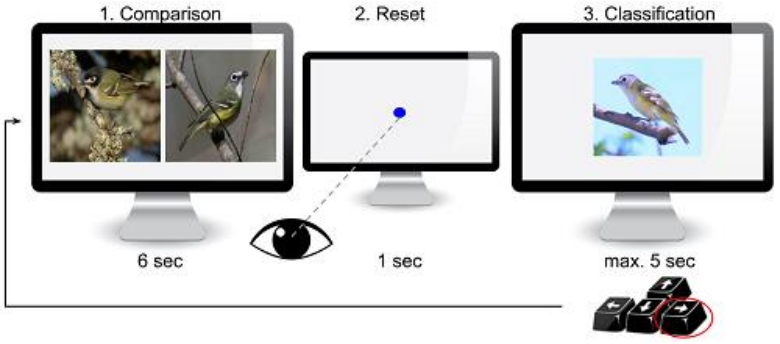
What particular example did you use in this work?

The main objective was comparing two types of birds, where we give the annotator six seconds to explore the objects and find the differences. Then we show one instance of the two previous classes, and give the annotator five seconds to make a decision to which class they think this image belongs. The five seconds was kind of short, and usually in gaze studies they use longer fixations. In our case it was a short time because the annotator had to give a fast reply, so we had to process the gaze data in a different way. We had to take out the outliers and have a shorter time for the fixations.



How did you solve the problem?

After we collected the gaze data, we processed the raw data and obtained the gaze points out of that. Then we wanted to come up with a class representation, which is needed in zero-shot learning to aid the classification task. To do this, we used the gaze data for the individual images of the class to get an image representation, and then averaged all these images to one class representation. So we had three types of representations, the details are all in the paper. We extracted many features from the gaze points - the location on the images, the duration, the pupil diameter of the annotator, and the sequence information between the points, which is the angle between subsequent points. Using this information we noticed that the location, duration and sequence was more helpful than using the pupil diameter. Studies say that pupil diameter helps indicating the concentration level of the annotator, but apparently as the annotator became familiar with the categories, their concentration dropped. So it wasn't very helpful to use this information, and we had better performance using only the other features.



What are the next steps for your work?

In future work we want to explore how to combine the gaze information from different images to represent one class in a better way. We would also like to do more experiments on more datasets. Our work compared species level of birds and pets, but we could explore more or larger fine grained datasets, for example asking how we can compare on the subspecies level.

You seem very passionate about this subject. What do you particularly like?

I always had this interest in computer vision and how vision works in humans, and how we understand that one object is different than another, just by one glimpse. So it was interesting for me to study the human behavior. Zero-shot learning is particularly interesting because as humans this is easy. For

example if I describe a giraffe to you in words, you would know how a giraffe looks without ever seeing one before, just by me telling you that it has a long neck and brown/yellow color. And then when you see it you are able to recognize it. But this is not the case with computers, and it's very interesting to be able to somehow transfer this knowledge to computers.

Does the fact that you are working with this subject change the way you look at the eyes of people?

It kind of affected my way of looking at people, because I was always wondering what the trigger is that we use when gazing at objects, and what makes us recognize them fast - the visual system is very fast. So this work changed my view on humans and how they focus on different regions. When I had to go through the data which we collected from 5 participants, I saw that they have different focus regions. For example one would only focus on the head always, or the body, and so you see that there is bias of the participant data.

"Eyes are the most important feature in the human face. The way that you look at people is a very strong way of communication"

Did you also try it yourself?

Yes, I did it myself - it took a long time, because I had to do all of them. I learned a lot about different birds species and dogs and cats, it was interesting. I tried it myself to make sure it's comfortable for others to do the experiment.

Is this something that interests you also before you started you scientific studies, the way that people watch?

Actually for me personally I always found that the eyes are the most important feature in the human face. The way that you look at people is a very strong way of communication.

How do I look?

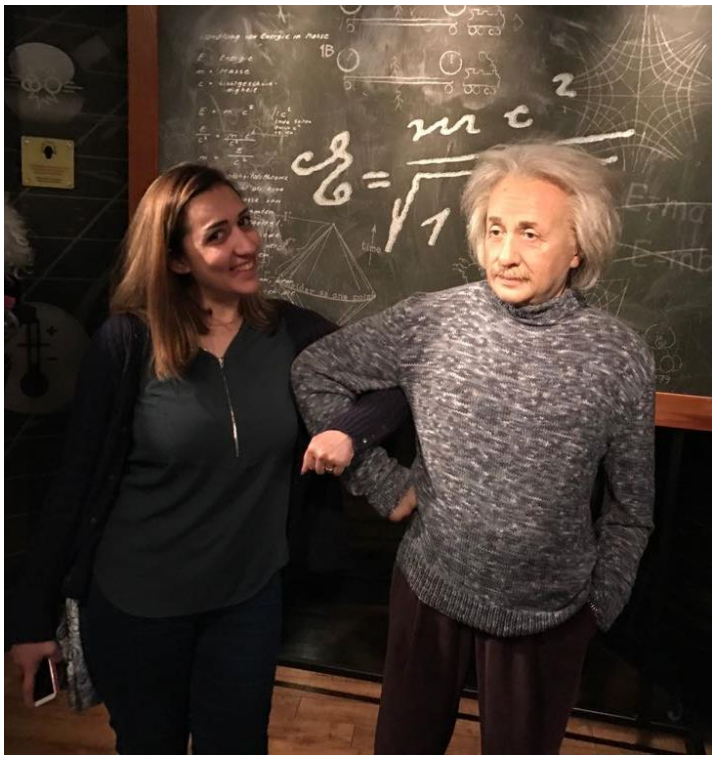
You look perfectly! [*We both laugh*]

But you know what the person is thinking of you, or if they like you, by the way they look at you, right?

You can obtain a lot of information by looking at the way they look at you.

How come that babies naturally look into your eyes - and not into your ears, for example?

It could be because they are always moving and thus attracting the attention. And they are also just nice to look at and sparkling [*she laughs*].



You have a special story: of the 5,000 CVPR participants, you are maybe the only one who lived in a war zone only three years ago. Can we tell our readers where you come from?

I originally come from Damascus, the capital of Syria, where I was born and raised. I came to Germany in 2014 to do a master's degree, and I already had plans to continue studying abroad. Germany especially has been very advanced in computer science recently. But of course, the war situation was the main motivation to leave the country.

"In other places that are not a war zone, you can build dreams without much worrying about the basic things"

You underwent many difficult things and had very strong experiences. It must be very difficult being a young woman in a war zone, dreaming of going away?

You're always feeling uncertain about everything. Whatever plan you come up with, you are always uncertain whether it will happen or not. In other places that are not a war zone, you can build dreams without much worrying about the basic things. But in the war zone, for example the electricity is unstable or the water station is unstable. So you are more focused on the basic life needs instead of focusing on your dreams.

Were you also worried about food?

In Damascus, especially in the city center, it was a bit better than in other places. But in the rural areas, it was sometimes under siege, and it was always hard for the people to get food.

Did these experiences make you stronger or weaker?

I think it made me stronger. Because I now know that as a human we are surprisingly adaptive to any situation. At the beginning, you will have a lot of fear and concern about even going out of your house. But then you grow a resistance and you are stronger to face these fears and just be able to continue your life. And I think I am now much stronger, because I know how to face my fears.

What is the latest challenge you had to overcome?

When I applied here, I wasn't sure if I would get the Visa or not, because of my Syrian passport. And even though I got the Visa, I was worried all the way here until I passed the borders, because I was unsure whether they would let me through.

I think that CVPR 2017 would have lost a lot if you wouldn't have had this Visa!

"I am now much stronger, because I know how to face my fears"

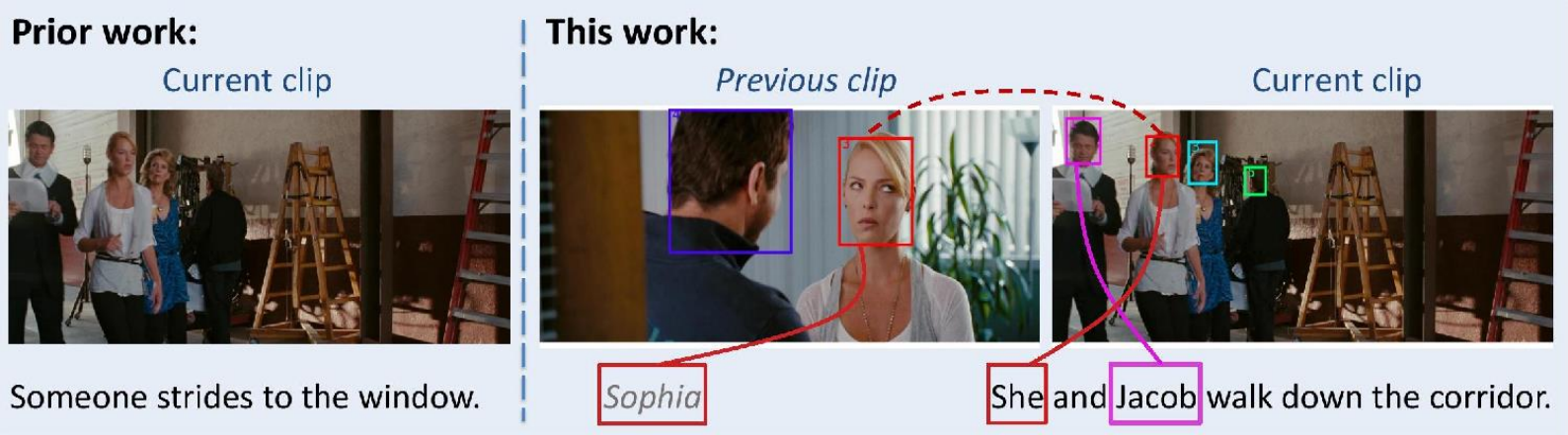
Generating Descriptions With Grounded and CoReferenced People

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



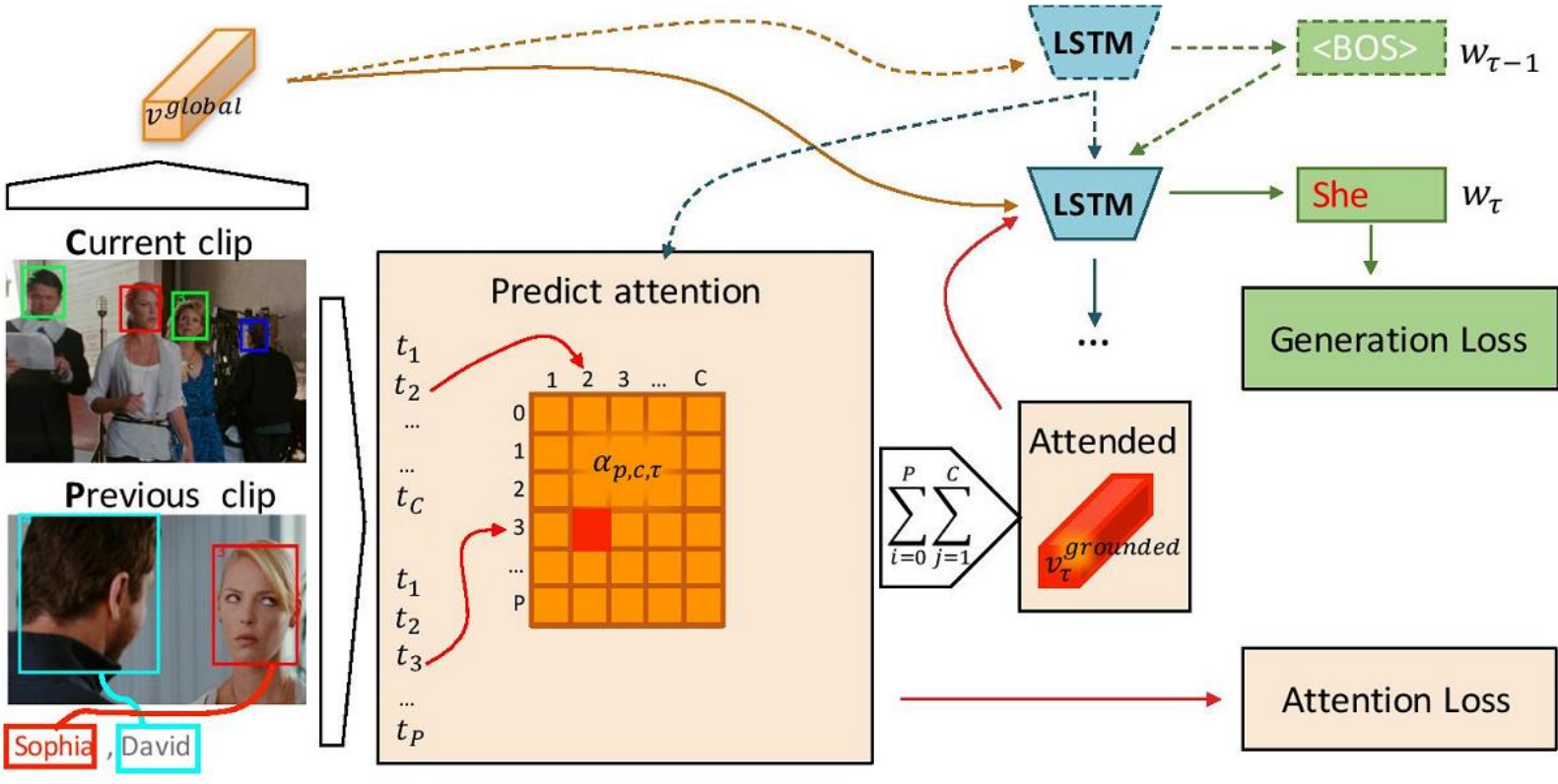
"Shoot for the moon; even if you miss, you'll land among the stars!"

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*".

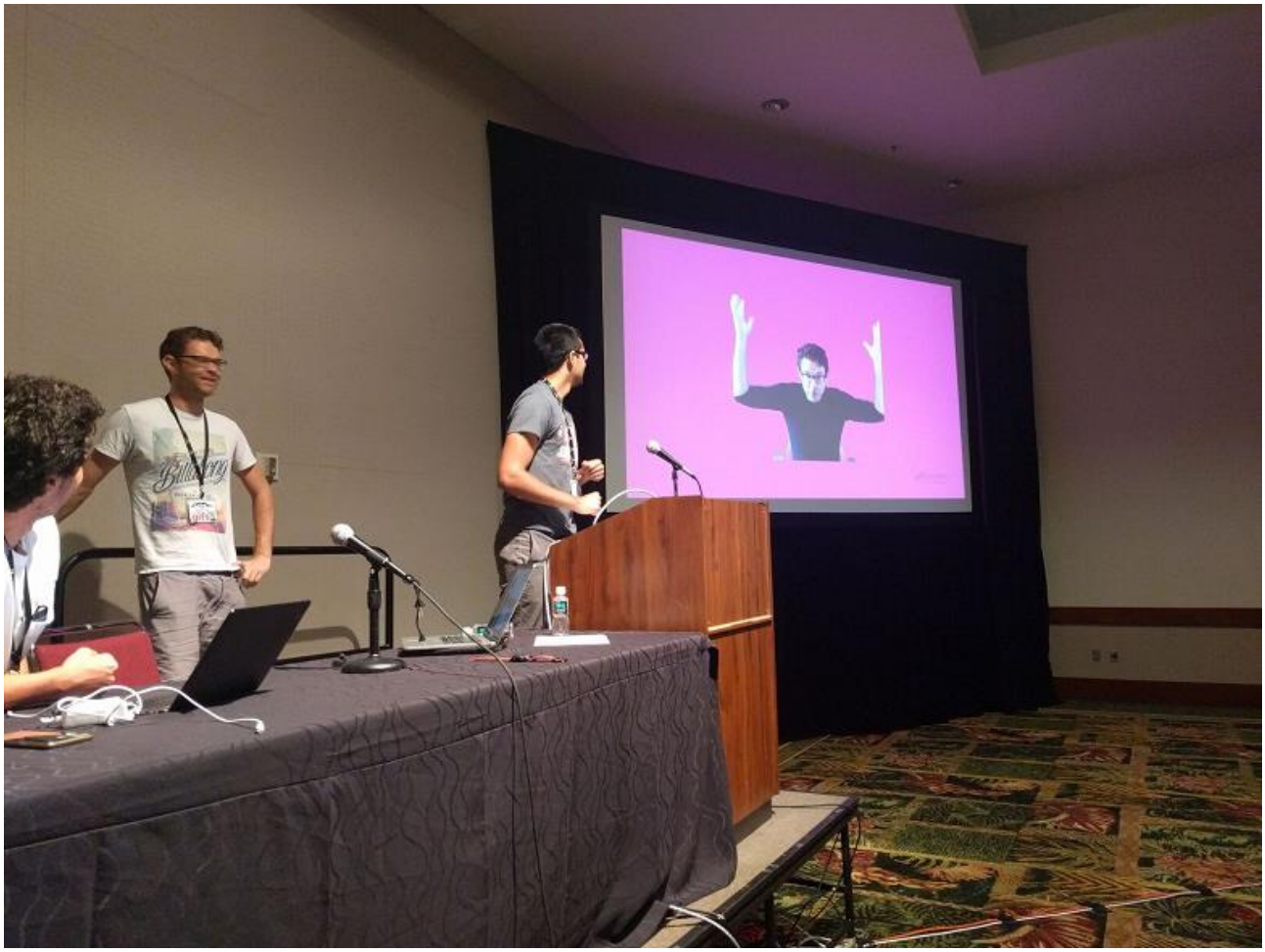


Densely Annotated Video Object Segmentation

by Jordi Pont-Tuset

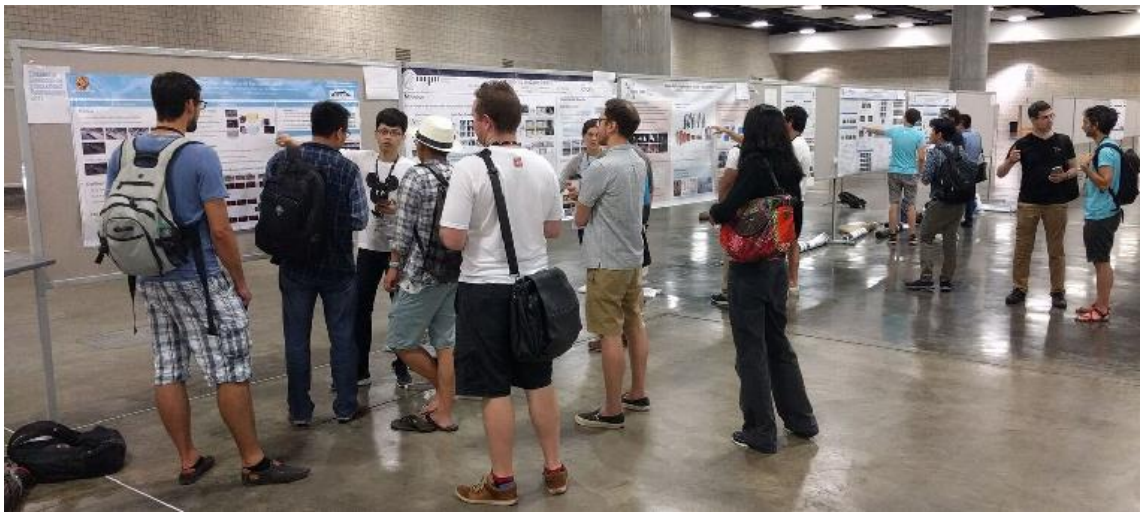
Every month, Computer Vision News reviews a **challenge** related to our field. If you do not take part in challenges, but are interested to know the new methods proposed by the scientific community to solve them, this section is for you. This month we dedicate this space to the DAVIS workshop and challenge, organized as a satellite event of CVPR2017. We published here a preview of the event and now Jordi Pont-Tuset tells you what actually happened at the workshop.

The 2017 **DAVIS Challenge** on Video Object Segmentation workshop (davischallenge.org) was held on the morning of the last day of CVPR2017. Video object segmentation is about separating the pixels of an object from the background in a video, and in particular, the challenge is about the semi-supervised approach, where the segmentation of the first frame is given and it should be propagated to the rest of the video.



The DAVIS dataset consists of **150 high-resolution video sequences**, annotated with pixel-level accurate segmentations, totalling more than **10.000 annotated frames**. The participants were given a subset of the annotations and they had the submit the results for the rest of the subset, for which the annotations are not public. The result could be submitted to an evaluation server that returned the quality of the result.

The best entries of the challenge presented their work in form of talks and posters, which included a variety of approaches and techniques. **Prof. Chenliang Xu** and **Prof. Fuxin Li** were the invited speakers from academia presenting their research; **R. O'Reilly** and **Dr. M. Gygli** from gifs.com explained how they use video object segmentation in real-world applications. There was a push towards publishing the code of the techniques, sharing results and the 2018 edition of the challenge was promised.



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.

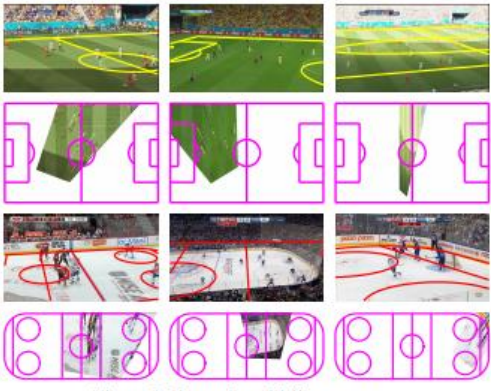


Figure 8: Examples of failure cases

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.

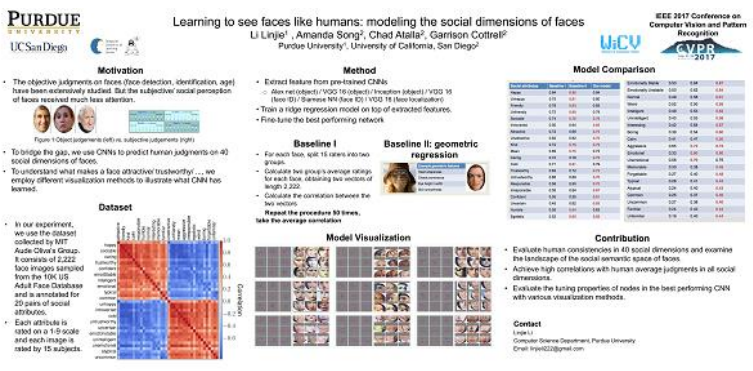
Namdar's supervisors and co-authors, **Raquel Urtasun** and **Sanja Fidler** (see pages 10-14 of this magazine) followed attentively our discussion; when asked to mention the main attractiveness of this work, Raquel told us that "*the key of this work is to come up with a parameterisation of the problem 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*".

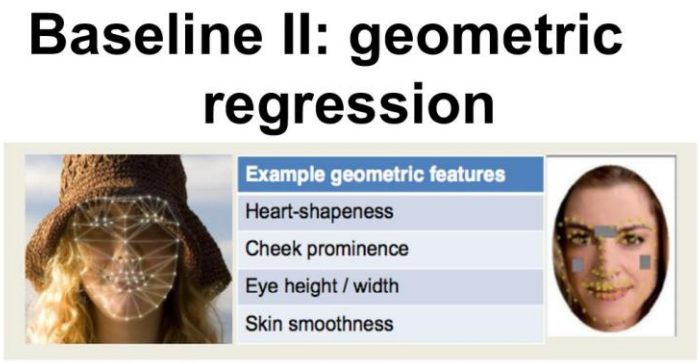
Learning to see faces like humans: modeling the social dimensions of faces

Linjie Li is currently pursuing her PhD in Computer Science at **Purdue University**. Prior to her Ph.D., she obtained a master's degree in Electrical Engineering from **UCSD**. She was working as a research assistant at **GURU lab** in UCSD, focusing on machine learning, computer vision and neural networks.

In the era of the digital age, we are constantly forming first impressions on others by browsing each other's photos online. Although first impressions seem to be subjective, psychological studies have shown that there is often a consensus among human in how they perceive attractiveness, trustworthiness, and dominance in faces. Are deep learning models, which have successfully conquered various visual tasks, also capable of predicting subjective social impressions of faces? To answer this question, we systematically examine 40 social features on faces and use deep learning models to predict human first impression on faces. Employing the internal representations from pretrained neural networks (for object classification, face identification, face landmark detection), we build a ridge regression model on top of the extracted features and our model can successfully predict human social perception whenever human have consensus. We further visualise the key features defining different social attributes to facilitate an understanding of what makes a faces salient in a certain social dimension.

This work, prepared with [Amanda Song](), **Garrison Cottrell** and **Chad Atalla**, was presented at the **Women in Computer Vision** (WicV) Workshop of CVPR2017.





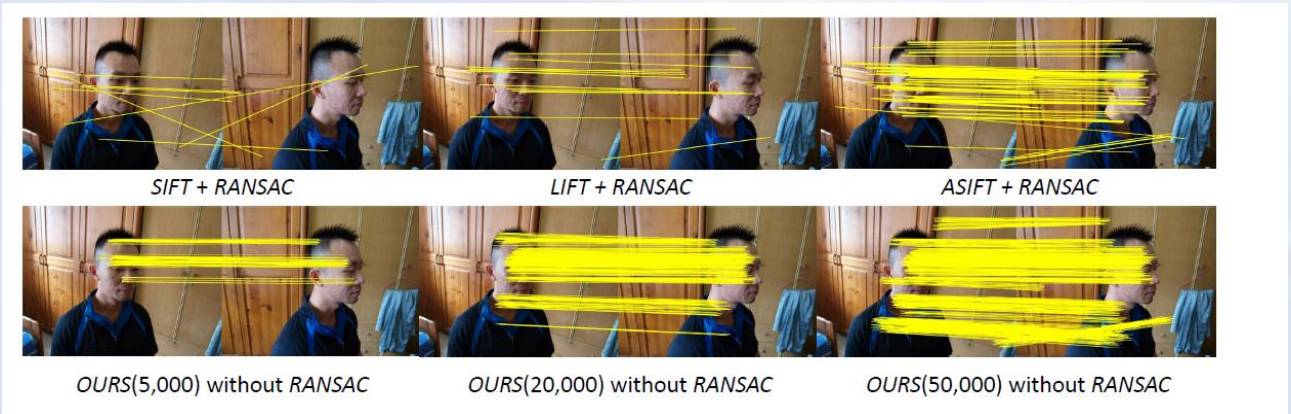
GMS: Grid-based Motion Statistics

Every month, Computer Vision News reviews a research from our field. This month we have chosen to review two papers. The first one is **GMS: Grid-based Motion Statistics for Fast, Ultra-robust Feature Correspondence**. a paperoffering a simple means to incorporate motionsmoothness in a way that rapidly and reliablydifferentiates true and false matches in a region. Weare indebted to the authors (**JiaWang Bian** and **Wen-Yan Lin** - as joint first authors - as well as **Yasuyuki Matsushita**, **Sai-Kit Yeung**, **Tan-DatNguyen** and **Ming-Ming Cheng**) for allowing us to use their images to illustrate this review. The website of the project is here.

"GMS achieved real-time matching of features in challenging scenarios, not yet successfully dealt with before"

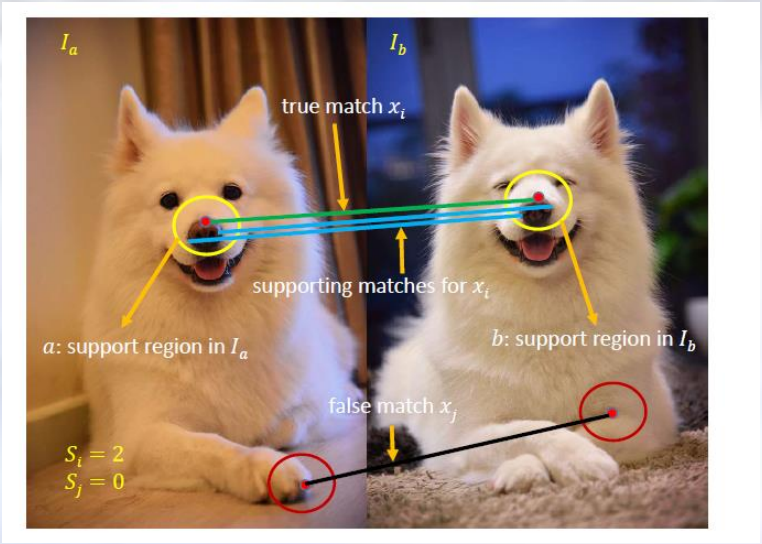
Background, motivation and novelty:

The main advantage of encapsulating smoothness constraints into feature matching is the ultra-robustness of the generated results. However, the price to pay is not negligible, in terms of complexity and slowness, which prevent the use of such a technique for video. The challenge is to develop accurate and robust matching of features between images, quickly enough and with computational efficiency sufficient for real-time application. GMS, that is Grid-based Motion Statistics, is the solution proposed by this paper to incorporate motion smoothness as a statistical likelihood of having a certain number of feature matches between a region pair by the way of differentiating true and false matches - by evaluating the number of matches in selected neighborhood. The assumption (largely inherited from previous works) behind this idea is that motion smoothness induces correspondence clusters that are unlikely to occur at random. Statistical measures (building on the law-of-large-numbers) are so introduced to reject false matches and allow previously unthought-of results.



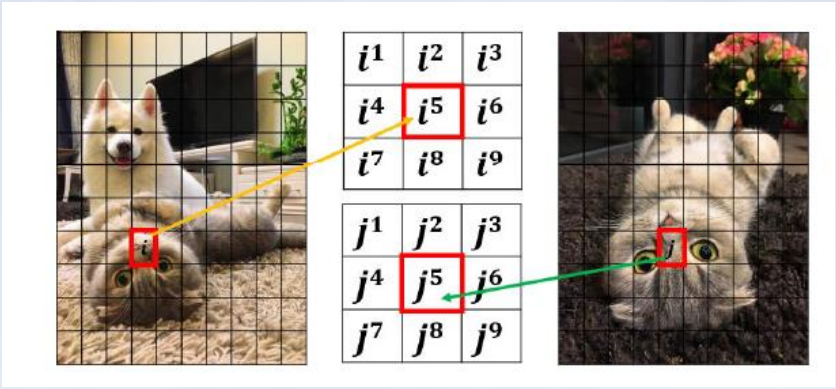
To summarize, the main advantages and contributions of GSM are: (1) Improved real-time matching based on rejection of false matches, by operationalizing statistical-based motion smoothness properties. (2) Real-time grid-based approach. (3) approach manages to match features in previously intractable scenes. (4) Significantly better performance than traditional SIFT, SURF and LIFT [see the September 2016 issue of Computer Vision News].

Given a pair of images taken from different views of the same 3D scene, a feature correspondence implies a pixel in one image is identified as the same point in the other image. If the motion is smooth, neighboring pixels and features move together. This allows to make the following assumption that motion smoothness will ensure that a tight neighborhood around a true match is the same 3D location in both images, while the tight neighborhood around a false match are two different 3D locations. Since true match neighborhoods are the same location -- there should be many similar features in the surrounding area (illustrate with yellow circles images below) which means many supporting matches in the neighborhood. Conversely, false matches should have significantly fewer supporting matches (red circles in both images). The authors operationalize this idea into the real-time matching algorithm described next.



Method:

Before diving into the algorithm let's define the notation we'll be using: Image pair $\{I_a, I_b\}$ have $\{N, M\}$ features respectively; $\chi = \{\chi_1, \chi_2 \ldots \chi_i \ldots \chi_N\}$ is the set of all nearest-neighbor feature matches between I_a and I_b. Images I_a and I_b. are both divided into G=(20x20) cell grids as illustrated above. We enumerate the cells of image I_a as i^k and the cells of image I_b asj^k. $|\chi_{i^k\ j^k}|$ is the number of matches between cells $\{i^k, j^k\}$.



S_{ij} is the score for cell-pair $\{i, j\}$. Defined by the equation $S_{ij} = \sum_{k=1}^{9} \left|\chi_{i^k j^k}\right|$. s_f (small s) denote the standard deviation of the binomial distribution of the false matches. The threshold τ approximated as $\alpha \cdot s_f$, which can in turn be approximated as $\alpha \cdot \sqrt{n_i}$. This last approximation gives us a per-cell threshold, defined as $\tau_i = \alpha \cdot \sqrt{n_i}$, where $\alpha = 6$ was empirically selected and n_i is the total number of features in the 9-cell neighborhood, as illustrated above.

Now we are equipped with all notation let's dive into the algorithm. The pseudocode is followed by explanations of the steps.

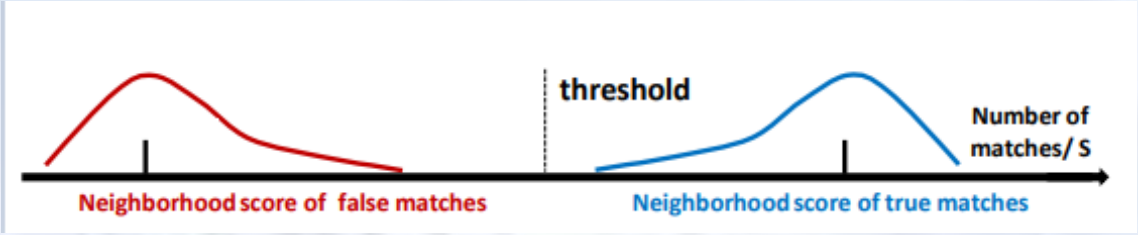
Input: One pair of images
Initialization:
1: Detect feature points and calculate their descriptors
2: For each feature in I_b, find its nearest neighbor in I_a
3: Divide two images by G grids respectively
4: **for** $i = 1$ to G **do**
5: $j = 1$;
6: **for** $k = 1$ to G **do**
7: **if** $|\mathcal{X}_{ik}| > |\mathcal{X}_{ij}|$ **then**
8: $j = k$;
9: **end if**
10: **end for**
11: Compute $\mathcal{S}_{ij}, \tau_i$;
12: **if** $\mathcal{S}_{ij} > \tau_i$ **then**
13: $Inliers = Inliers \cup \mathcal{X}_{ij}$;
14: **end if**
15: **end for**
Iteration: Repeat from line 4, with gird patterns shifted by half cell-width in the x, y and both x and y directions.
Output: $Inliers$

The input of the algorithm is a pair of images and the output is the matching features denoted as *Inliers*.

The algorithm starts by detecting ORB features in both images and divided them into G-cell grids (lines 1-3). Next, the algorithm loops over the cells in one image and finds the cells with the highest number of matching features from the other

image (lines 4-10). S_{ij} and τ_i are computed for the 9-cell neighborhood of the cell with the highest number of matching features for the i-th iteration (line 11). If the number of matching features exceeds the threshold, those features are included in the result (*Inliers*).

The threshold condition eliminating false matches (line 12) is based on the statistical model of the different distribution of supporting matches in the surrounding region in the case of true and false matches.



Readers who want to read more about the model and the statistical assumptions can find in depth details in section 2 of the article.

Evaluation and Results:

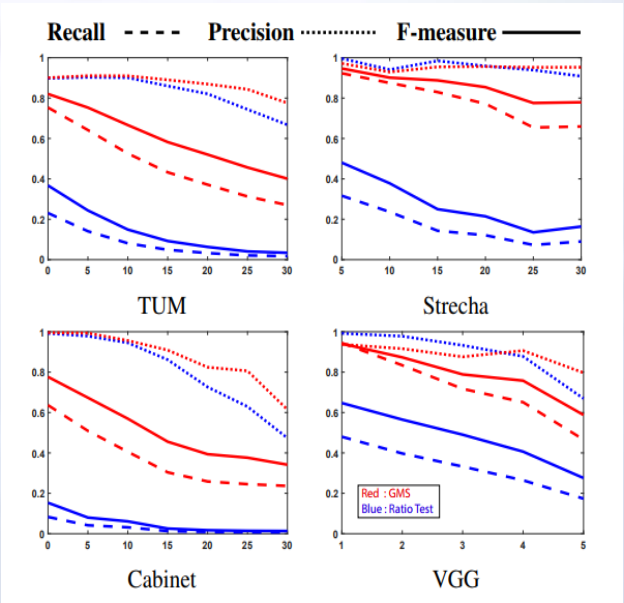
The paper used four datasets detailed in the table below:

Dataset	*TUM*	*Strecha*	*VGG*	*Cabinet*
Full name	RGB-D SLAM Dataset and Benchmark	Dense Multiview Stereo Dataset	Affine Covariant Regions Datasets	A subset of TUM dataset
Image pairs	3141	500	40	578
Ground truth	Camera pose, Depth	Camera pose, 3D model	Homography	Camera pose, Depth
Description	Including all image condition changes	Well-textured images	Viewpoint change, zoom+rotation,blur	Low-texture images with strong light

Results are divided according to several metrics:

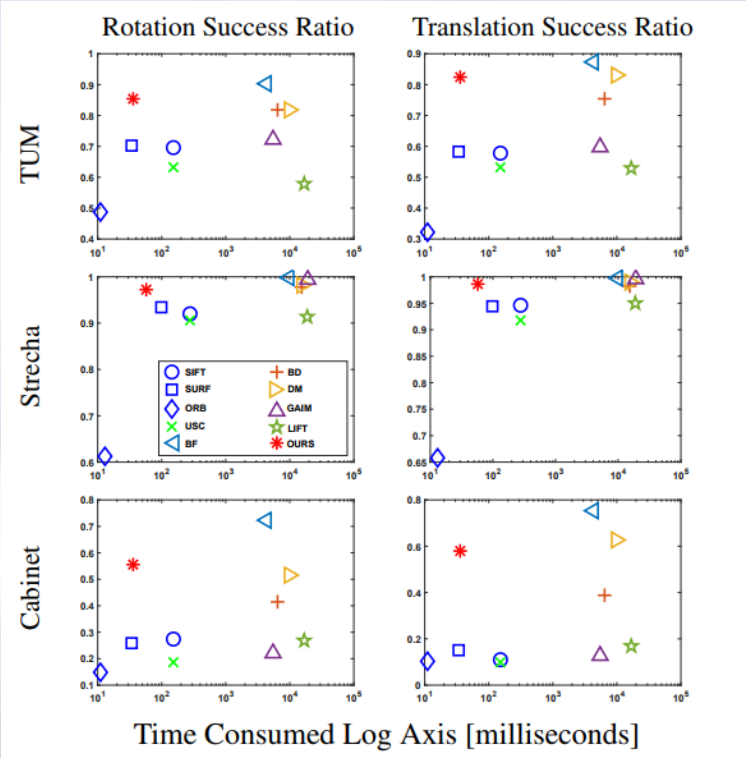
(1) Precision and Recall:

GMS was compared against ratio-test (with a threshold of 0.66). Precision (dotted lines) was very close, but in recall and F-measure GMS clearly outperformed



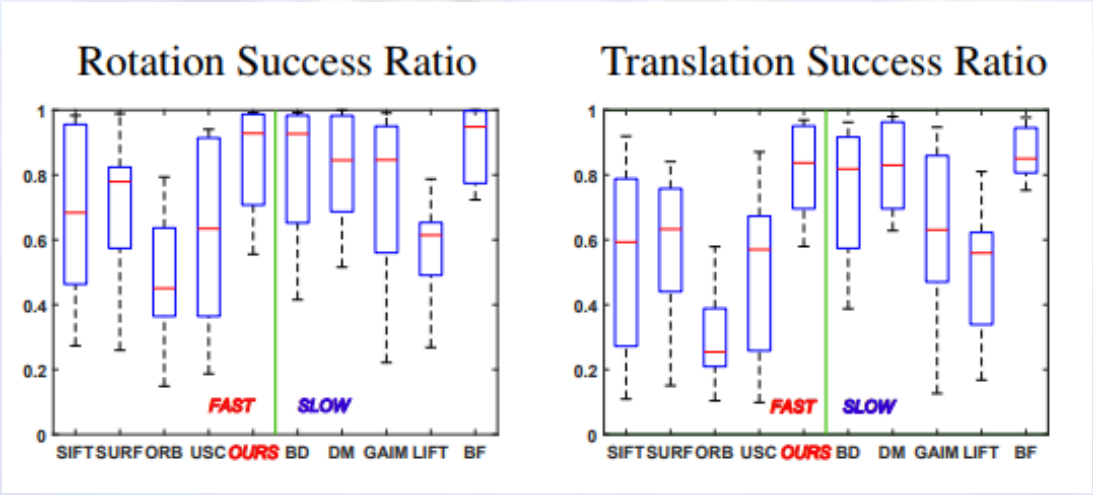
(2) Performance/Speed Tradeoff:

Compared to 9 other methods, GMS shows performance near that of the best, and speed consistently among the quickest. None of the other top-performing methods show speed near that of GMS. This can be particularly useful for real-time applications, where speed may be the more critical consideration in judging algorithms.



(3) Consistency:

The figure below illustrates performance variation across different TUM scenes. The red mark is the median, the blue box covers the 25th to 75th percentiles. Whiskers show performance beyond the 25th and 75th percentiles. GMS (the fifth from the left, labelled in red) is the most consistent fast algorithm, its consistency comparable to that of much slower ones.



Sum-up:

GMS achieved real-time matching of features in challenging scenarios, not yet successfully dealt with before. It shows high performance, with much better speeds than comparably performing feature extraction methods.

Computer Vision News lists some of the great stories that we have just found somewhere else. We share them with you, adding a short comment. Enjoy!

The Difference Between Deep Learning and Machine Learning

In the never ending debate over this question, we recommend you read Arthur Chan's opinion, which is as usual very meaningful and to the point. Read it with all its quotes and references in Arthur's excellent blog here.

Computer Vision Applications in Mental Health: Dr. LP Morency

When did you last hear about how our community's work can contribute to mental health? This is why we want you to read this passionating interview with this **Carnegie Mellon** professor working on these issues and suggesting algorithmic solutions to support clinicians in assessing **depression**, **anxiety** and other mental illness. Read More

Why Google Can't Tell You if it Will Be Dark When You Get Home

Or why Google has an answer for practically everything but a few apparently simple questions. **Emmanuel Mogenet**, head of **Google Research Europe**, explains why he and his team of 130 Googlers in **Zurich, Switzerland**, cannot get to the semantic meaning of this question and are desperately trying to change this. Read More...

AR Helmet Helps Firefighters See through Smoke and Get out of Fire

A new system called **C-THRU** provides real time navigation through a transparent AR display placed directly within the firefighters line of sight. It is developed by a company called **Qwake Technologies** using also principles from the field of visual neuroscience. It aims to help firefighters get in an out of the fire five times faster. Read More

Microsoft, Apple Patents Hint at Hand Gesturing in Home Appliances

Both **Microsoft** and **Apple** have files patents disclosing a strong interest in developing highly sophisticated next-gen **Natural User Interfaces** (NUIs): advanced hand gesturing systems that will one day work with home appliances, desktops, notebooks, vehicles and so forth. Read...

You will enjoy reading:

- MIT CSAIL Research Offers a Fully Automated Way to Peer Inside Neural Nets
- Computer Reads Body Language in Real-time: sees hand poses and tracks multiple people
- AI Vision Can Determine Why Neighborhoods Thrive

RSIP VISION | Global Leader in Computer Vision

JOIN THE AI AGE IN OPHTHALMIC IMAGING

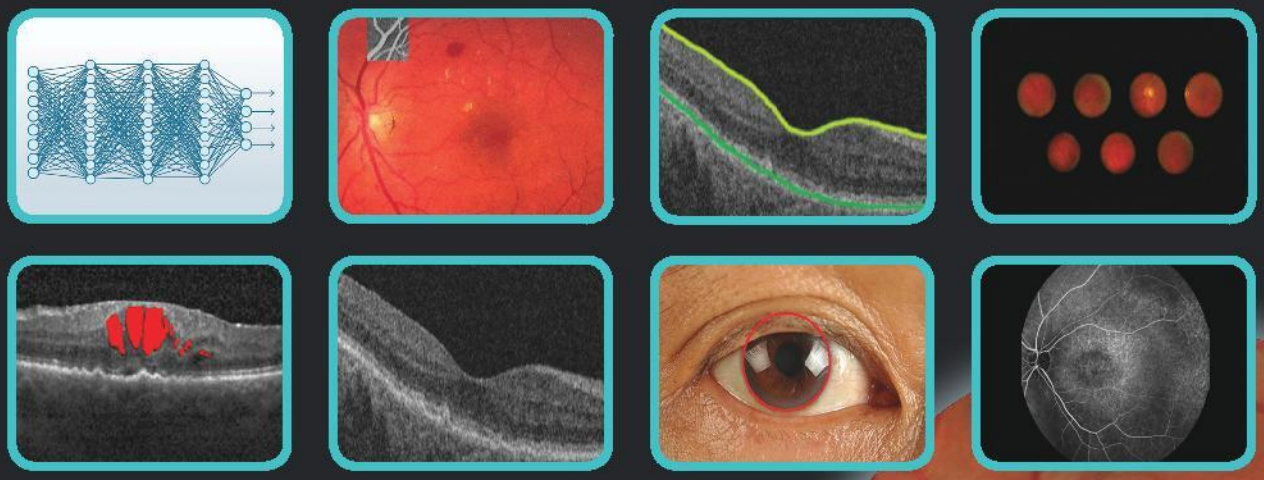
Be part of the AI revolution in ophthalmic imaging. RSIP Vision, a global leader in computer vision and image processing, has completed numerous ophthalmology projects involving development of advanced algorithmic software. In order to achieve state-of-the-art results, the company has invested in breakthrough technology like AI and deep learning.

Your applications can be first-class too! Talk with us about how to gain a competitive edge which will give your device a top-notch performance and a huge advantage in the marketplace. New technologies can benefit all eye care fields in all their components, from telemedicine to diagnosis, from data collection to cloud operations.

Consult our experts now!

Ophthalmology software Research & Development:

- Machine Learning
- Artificial Intelligence
- Image Processing
- Deep Learning
- Computer Vision
- Ophthalmic Imaging



Proud sponsor of ARVO The Association for Research in Vision and Ophthalmology

www.rsipvision.com

Upcoming Events

ECVP - European Conference on Visual Perception
Berlin, Germany August 27-31 Website and Registration

BMVC 2017 British Machine Vision Conference
Imperial College London, UK Sep 4-7 Website and Registration

MICCAI - Medical Image Computing & Computer Assisted Intervention
Quebec, Canada MEET US! Sep 10-14 Website and Registration

ICIP 2017 IEEE International Conference on Image Processing
Beijing, China Sep 17-20 Website and Registration

RE•WORK Deep Learning Summit
London, UK Sep 21-21 Website and Registration

SIPAIM - Int. Symp. on Medical Information Processing and Analysis
San Andres - Colombia Oct 5-7 Website and Registration

RE•WORK Deep Learning Summit
Montreal, Canada Oct 10-11 Website and Registration

Vipimage on Computational Vision and Medical Image Proc.
Porto, Portugal Oct 18-20 Website and Registration

ICCV 2017
Venezia, Italy Oct 22-29 Website and Registration

EMMCVPR 2017 - Energy Minimization in Computer Vision
Venezia, Italy Oct 30-Nov 1 Website and Registration

Did we miss an event?
Tell us: editor@ComputerVision.News

FREE SUBSCRIPTION

Dear reader,

Do you enjoy reading Computer Vision News? Would you like to receive it **for free in your mailbox** every month?

Subscription Form
(click here, it's free)

You will fill the Subscription Form in **less than 1 minute**. Join many others computer vision professionals and receive all issues of Computer Vision News as soon as we publish them. You can also read Computer Vision News on our website and find in our archive new and old issues as well.

We hate SPAM and promise to keep your email address safe, always.

FEEDBACK

Dear reader,

How do you like Computer Vision News? Did you enjoy reading it? Give us feedback here:

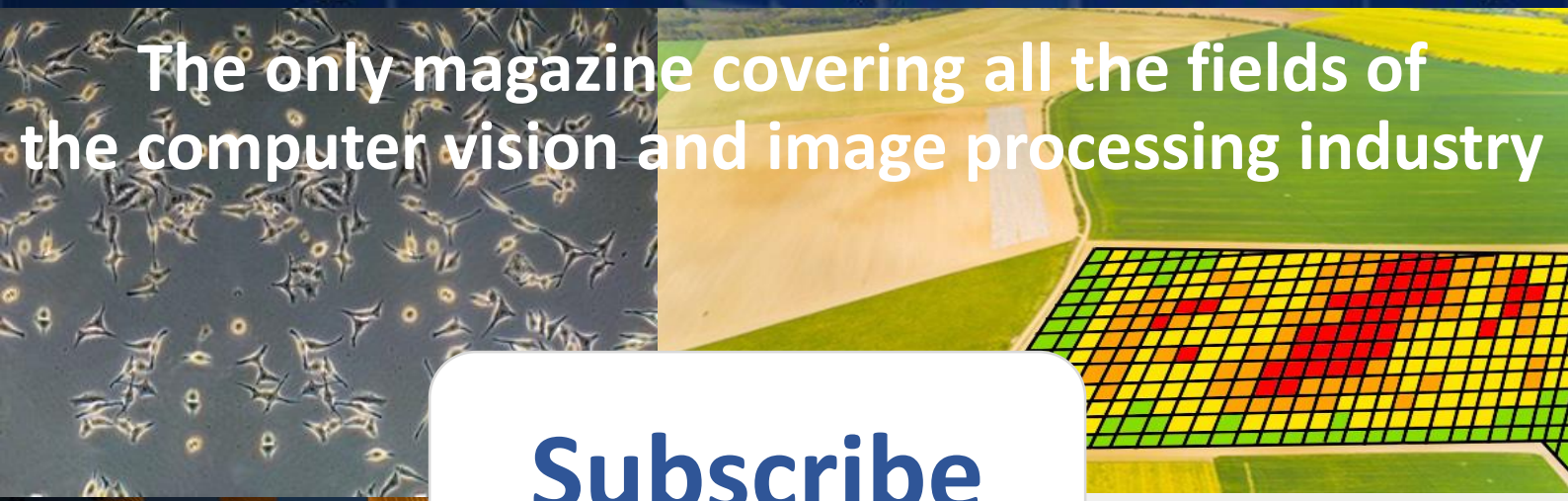
Give us feedback, please (click here)

It will take you only 2 minutes to fill and it will help us give the computer vision community the great magazine it deserves!

Improve your vision with

Computer Vision News

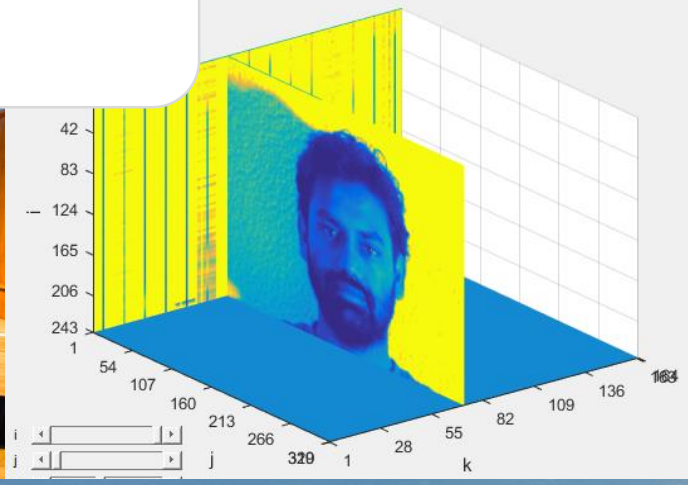
The Magazine Of The Algorithm Community



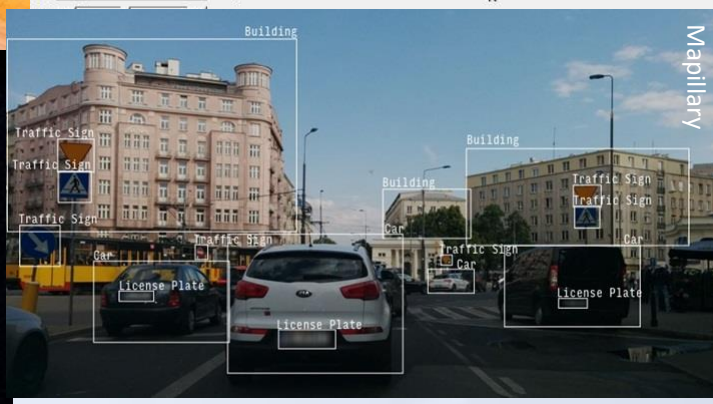
The only magazine covering all the fields of the computer vision and image processing industry

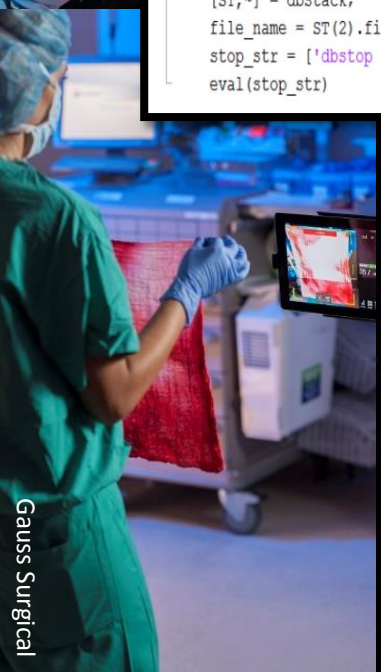
Subscribe

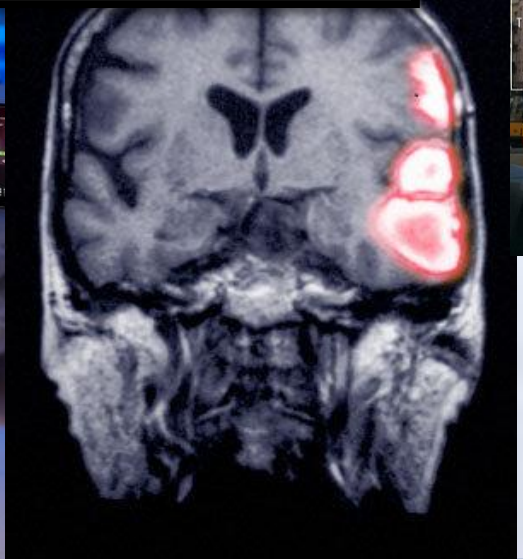
(click here, it's free)




```
% 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