Computer Vision News

The magazine of the algorithm community

April 2017

Guests:
Arthur Chan - Waikit Lau
AIDL - Artificial Intelligence and Deep Learning Facebook Group

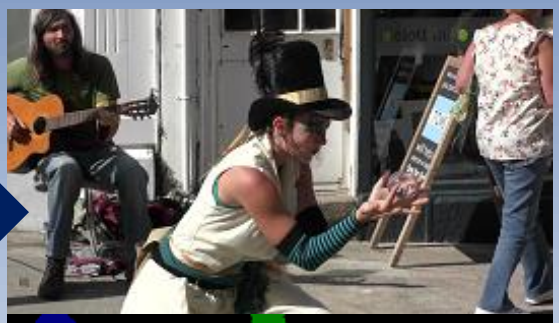
Research Paper:
YOLO9000: Better, Faster, Stronger Real-Time Object Detection

Project:
Navigation Systems for Orthopedic Surgery

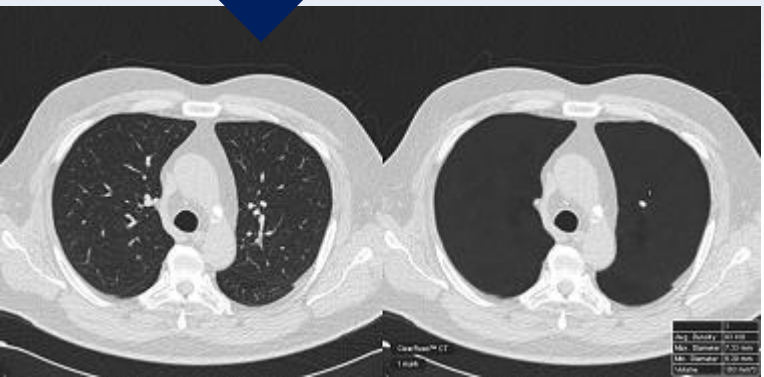
We Tried for You:
Cyclops: Video-conferencing with Computer Vision and AR

Women in Computer Vision: Tal Arbel

DAVIS Challenge
Video Object Segmentation



Application:
Riverain Technologies
Lung disease detection



Spotlight News

Tool:
SIFT3D - SIFT in 3D

Events:
Vision Monday
Leadership Summit

Project Management:
Computer Vision Software Projects

A publication by

RSIP VISION

Application of the Month
Riverain Technologies

04

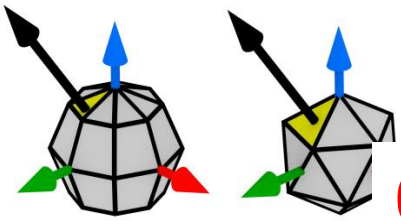
Women in Computer Vision
Tal Arbel

14

Project of the Month
Navigation Systems for Orthopedic Surgery

28

Tool of the Month
SIFT3D



08

Guests
Waikit Lau Arthur Chan Samson Timoner

19

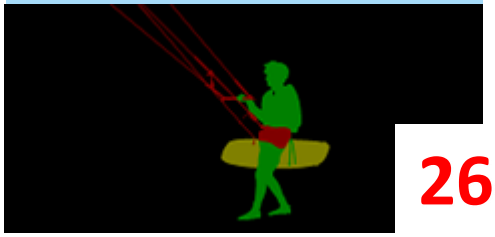
Research - YOLO9000
Real-Time Object Detection

34

Spotlight News

13

Challenge
Video Object Segmentation



26

Upcoming Events

41

Dear reader,

With this April issue of **Computer Vision News** we start the second year of publication: last month, we weighed the outcomes of this journey, and they are tremendously positive. Now it is the time to think about the plans for the future. We treasure your feedback: we will build on it to keep giving you fresh, proprietary and original readings, to fulfil the need for a community magazine which (according to what we hear) no other communication tool does. One major example is the on-site coverage of the major conferences, with which we partner to publish the **CVPR Daily**, the **ECCV Daily** and the **MICCAI Daily**. Another example is the **free subscription to our magazine** which is offered to all, including every member of your team and of your class.

Last month, we also promised you our **first special edition**: the challenge we took on ourselves was successful, with the publication of **Boston Vision**, a magazine that uncovers some of the vibrant vitality animating the image processing and computer vision community in Boston. When you finish reading Computer Vision News, follow the link and read **Boston Vision** too!

This Computer Vision News of April features (as always) plenty of unique content and great guests for you: you will be fascinated by the work of **Riverain Technologies** on lung disease; you will be captivated by the interview with **Waikit Lau** and **Arthur Chan**, founders of **Artificial Intelligence & Deep Learning**, the most successful Facebook group in our community; you will be enchanted by **Tal Arbel**, an impressive woman scientist doing great work with her students in both computer vision and medical image analysis. Don't miss the Tool, the Research, the Project, the Spotlight News, the Events and last but not least the Challenge, which this month has been specially written for you by **Jordi Pont-Tuset**.

We are confident that you will love reading this April magazine. Please continue sharing it with colleagues and friends.

Enjoy the reading!

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

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

Contact us

Give us feedback

Free subscription

Read previous magazines

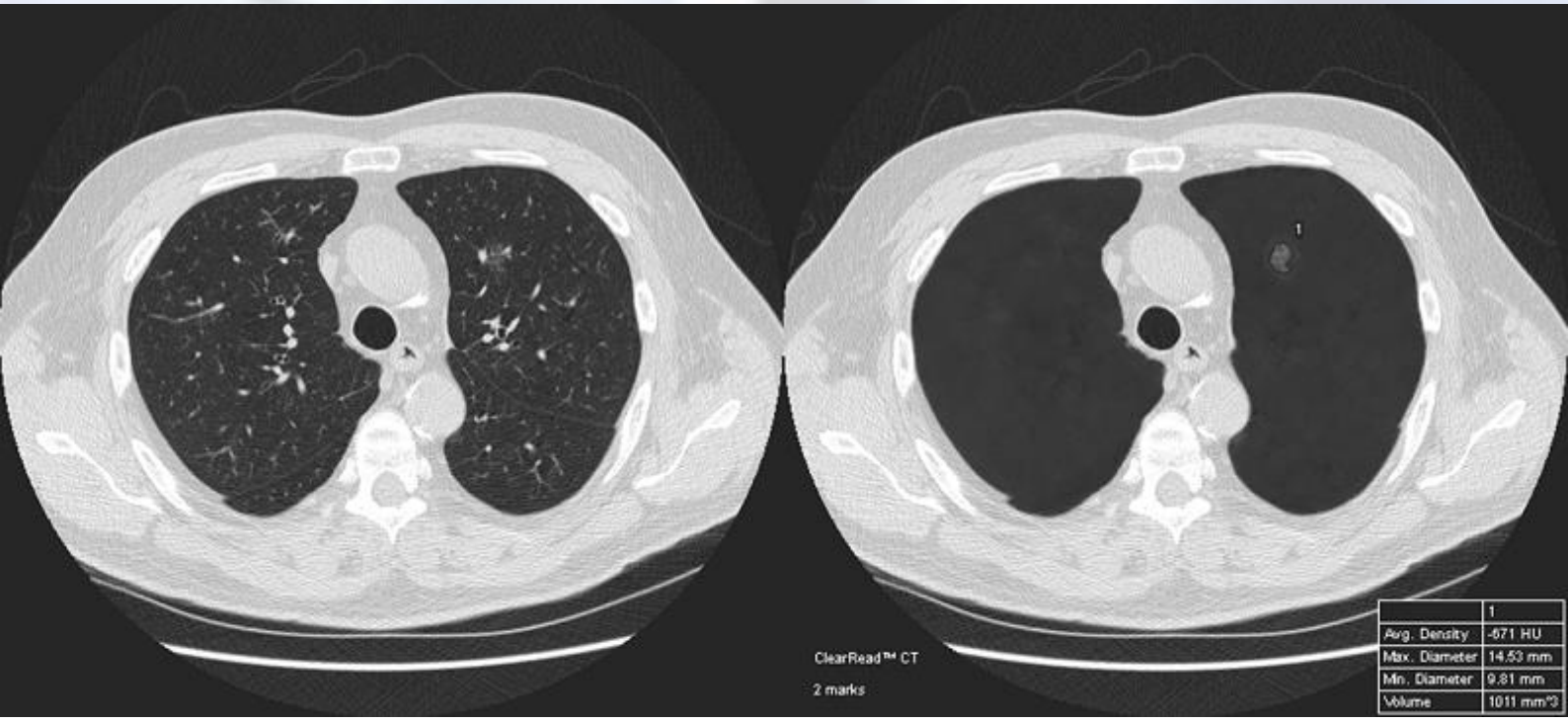
Riverain Technologies

Riverain Technologies provides software tools to aid clinicians in efficient and effective early detection of lung disease. With about five million deadly cases per year and survival rate being related to the stage in which the disease is diagnosed, early detection is a key priority. In addition to tools to facilitate early detection, Riverain tools allow faster reading, increasingly critical in radiology. We talked about it with CEO **Steve Worrell** and Chief Science Officer **Jason Knapp**. Riverain's core competencies are machine learning and image analysis. They are very much dedicated to medical imaging applications, more specifically **thoracic imaging in x-ray and CT**. They have four FDA cleared X-ray products and one FDA cleared CT product they offer to the marketplace.

"We use modeling extensively for constructing training data, which I think is quite unique"

The CT product they launched this past September uses **deep learning** as well as many other novel technologies. Their thoracic imaging solutions help radiologists read more efficiently (faster) and more accurately in the detection of lung disease; in particular, the CT product aims at aiding the radiologist in **detecting lung cancer** while allowing them to read faster.

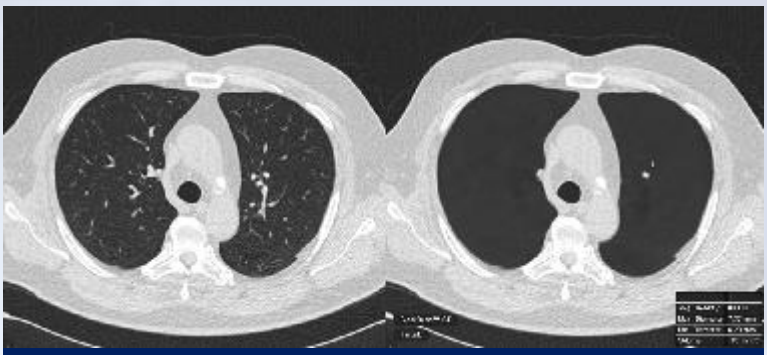
When asked about what makes their technology stand out, Worrell explains that their solutions have unique attributes regarding both deployment



ClearRead CT's output for a non-solid nodule: this and next page's image show a slice from the original CT series at left, while the corresponding slice from the vessel suppressed volume is at right. ClearRead CT locates, segments, and makes measurements for each region. It provides the radiologists with the vessel suppressed series and the automatic segmentation/measurements for concurrent review - that is, they can see content while looking at the original data upon their initial review. The company claims that to the best of their knowledge, this is the first automatic detection system approved to do so. This is why radiologists can read faster, while finding more suspicious regions.

and development: they use modeling (data synthesis) quite extensively as part of the development process. In his own words: "*In machine learning applications, you face the challenge of collecting and labeling data in order to* ***train algorithms*** *to recognize things of interest. One of the novel things that we do as part of the development process is, rather than rely exclusively on acquired data, we use modeling extensively for constructing training data, which is quite unique, I think, relative to other companies in the medical imaging space; that's how we go about the development process from a training perspective to ensure we have representative data across the spectrum of situations that can arise, whether in the context of the disease itself or the anatomy of the patient.*"

Their work involves other novelties deserving to be known. For instance, **image normalization technology** (or domain adaption, in the machine learning vernacular). One thing that they try to do across all of their products is to build a solution that scales across, not just one device, but all devices in a clinical setting. Hospitals may have multiple CT devices in one facility or across a health network, coming from different manufacturers. Furthermore, different sites might have different acquisition protocols. Through the use of image normalization technology, their products work seamlessly across acquisition devices and imaging protocols. This has a lot of practical value as it provides scalability and ease of installation. They can deploy their solutions across an entire site or (in some countries like the United States) across a large network of hospitals and health centers through the change of a license file.



ClearRead CT's outputs for a solid nodule

The CT product facilitates nodule detection by the reader. A significant impediment to **nodule detection in the lungs**, not just for machines but also for radiologists, is represented by the vascular structures in the lungs. One of the unique aspects of their product is that it suppresses the vascular structures with the objective of making lesions more conspicuous and easier to detect by the radiologist. They estimate, for each voxel in the 3D CT scan, what fraction is occupied by a vascular structure versus a nodular structure. It's a continuous real value prediction, forgoing the use of explicit segmentation techniques.

"What you really want is the unusual cases, which can unfortunately be hard to come by"

Worrell comments: "*Through this process, we can achieve the objective of suppressing vascular structures and have it be much more robust than we would if explicitly segmented the vascular structures, which can be a fairly noisy process.*"

When you ask Knapp about the **algorithmic techniques** used to achieve this task, he jokes: "*Well, surprisingly, we've found deep learning quite useful. Seriously, though, as Steve suggested, we use a lot of modeling and data synthesis. We generally form a type of forward process, and use deep learning*

to learn the inverse process. This design principle is applied again and again throughout our applications. ***It is remarkably powerful****."*

"It's a very exciting time to be working in the field!"

He then adds: *"What some people new to the medical domain are finding out is that models built in a lab setting might not work well in a clinical setting due to the wide range of conditions that exist clinically. This is not a result of poor technique, but a consequence of using models driven by statistical correlations - and acquisition processes, as one example, can have a substantial effect on the nature of these correlations. Dealing with acquisition may not sound very exciting, but it is critical if you want to build clinically viable systems. Successful development really comes down to these types of fundamental issues."*

Knapp explains that the one problem is simply statistical size versus sample size. When people collect data sets they really focus on sample size, not statistical size. *"What you really want is the unusual cases, which can unfortunately be hard to come by. This is one of the hardest aspects about medical data: it has* ***a really long tail!****"*

To that Worrell adds: *"That's what radiologists tend to be really good at: extrapolating to those scenarios that are statistically infrequent, but nonetheless very important from a patient health standpoint."*

And Knapp: *"I think this is all tied together. Radiologists draw upon a tremendous amount of abstract knowledge when interpreting medical images, which allows them to reason in low SNR situations, handle ambiguity effortlessly, utilize causality, and learn from very few examples. And it is not at all clear how much of this knowledge can be learned just from annotated volumes. For tasks that are largely just pattern recognition, like* ***finding tumors or segmenting organs****, I think we [medical image analysis community] are in good shape, but there is obviously a lot of work to be done. It's a very exciting time to be working in the field!"*

An image of those directly involved in the algorithm work (from left to right): Praveen Kakumanu, Steve Worrell, Jason Knapp.

Worrell concludes: *"What I personally think is interesting with the products that we're developing is how the user interacts with the products and how the user perceives the product.* ***The dynamic between the machine and the human*** *is quite interesting. One of the things that Jason and I have learned is that it only takes a couple of bad results to totally turn a user against a machine solution:* ***robustness and predictability*** *of the machine are critically important."*

A publication by

RSIP VISION

Boston Vision

The vibrant Boston algorithm community

A special edition of Computer Vision News

Vision Systems:
Intelligent, automatic and robust computer vision technologies

Realeyes:
Measure people's reactions to something that they see

StopLift:
Video analytics and computer vision for POS revenue assurance

Quanterix:
Detect and measure molecules at very low concentrations

Affectiva:
Bringing emotional intelligence into the digital world

Computer Vision in Boston:
20 pages - click here to read!

Teledyne DALSA:
Software & hardware for machine vision applications


Lydey Automation

MERL:
Electronics & Communications, Multimedia, Data Analytics, Computer Vision, Mechatronics
"With computer vision as a hammer, every product looks like a nail!"

Cognex:
Advanced machine vision and industrial barcode reader systems

March 2017

SIFT3D

SIFT3D is a 3D image extrapolation of **SIFT (scale-invariant feature transform)**. It extracts a robust description of the content of 3D images by leveraging volumetric data to detect keypoints. It can also be used for 3D image registration with the **RANSAC algorithm**, by matching SIFT3D features and fitting geometric transformations.

SIFT3D is an open-source code, implemented C, and includes **Matlab** (Mex) wrappers.

Identification and extraction of local keypoints is a broad, well-established and popular area of computer vision. One of the best known and most popular methods is SIFT. Techniques for feature extraction (and SIFT in particular) were developed with 2D images in mind and cannot be trivially expanded to 3D images. SIFT3D is a robust tool for extracting a local keypoint description from 3D images; and is particularly useful and relevant in the medical domain for imaging techniques such as CT and MRI.

In the original 2D SIFT algorithm, described by Lowe in his paper "**Distinctive Image Features from Scale-Invariant Keypoints**", he broadly separated the process into three main stages:

1. **Scale-space detection and keypoint localization:** identifies points of interest (which the algorithm terms **keypoints**) that are independent of scale or location within the image. And a stable parabolic model is fitted to each keypoint.
2. **Orientation assignment:** assigns one or more orientations to each keypoint. This orientation assignment makes the points invariant with respect to rotation and viewpoint.
3. **Keypoint descriptor:** here, a gradient for each selected keypoint in each scale is computed. The gradients are transformed into a representation using a histogram, which enables their quantification and comparison relative to other keypoints.

Next, a detailed explanation of each of the three stages will be given, together with means used to expand and adapt the algorithm to enable SIFT3D to handle three-dimensional images.

1. **Scale-space detection and keypoint localization:** The first stage searches for potential points of interest, over all image locations and scales. The keypoints are derived in an efficient manner, identical to that of 2D SIFT: first, different Gaussian scalespace (GSS) levels are calculated by a Gaussian function with changing variance. Then, the Difference of Gaussians (DoG) is calculated by subtracting consecutive GSS levels, arriving at the local gradient per scale, which we denote as d for that point. The potential keypoints are denoted by $d(x,\sigma)$, where σ represents the scalespace level and x represents the location. Each point whose $d(x,\sigma)$ value exceeds the predetermined threshold

α is selected as a keypoint. Finally, as in 2D SIFT, we fit a parabolic model to each keypoint and the model is constructed by interpolating neighboring keypoints.

2. **Orientation assignment:** The second stage assigns an orientation to each keypoint derived at stage 1. The orientation for each keypoint is assigned based on image gradient directions. From here on out all operations are performed on keypoints that have undergone transformation based on their location and scale. This process produces keypoints invariant to image rotation.

In **two-dimensional SIFT** the orientation angle θ at any keypoint is determined by the gradients around that keypoint, and the rotation matrix R(θ) is derived from θ. However, the orientation is more difficult to arrive at in 3D.

There are several approaches to determining 3D orientation, SIFT3D implements one relatively straightforward approach, calculating the correlation between gradient components in the image using the following formula, referred to as the structure tensor:

$$K = \sum_{x \in W} \omega_x \nabla I_x \nabla I_x^T$$

where ∇I_x is the derivative of the image I at location x. And ω_x is the Gaussian weight component for window width W. This structure tensor is real and symmetric, and thus it has an orthogonal eigen-decomposition:

$$K = Q \Lambda Q^T$$

Matrix Q still doesn't give us an unequivocal orientation, for it is ambiguous as to the direction of change along each axis. To avoid this ambiguity, SIFT3D adds the condition that the derivatives along each axis be positive. For this, each column r_i of the rotation matrix R is calculated as follows:

$$r_i = \mathrm{sgn}\left(q_i^T d\right)$$

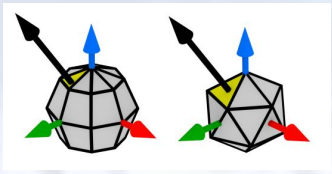
$$d = \sum_{x \in W} w_x \nabla I_x$$

where q_i is the i-th column of Q.

At this point, after the gradients have been calculated for each keypoint at the selected scale, they are transformed into gradient histogram representation to allow for their quantification and comparison relative to other keypoints.

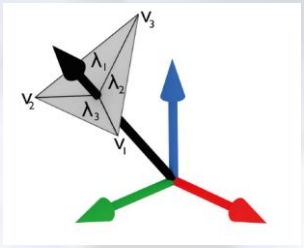
The gradient histograms represent the distribution of gradient directions in the window around the keypoint. Gradient histograms are considered a stable, reliable representation.

In 2D SIFT a 10-bin histogram was constructed by dividing the anglespace into 10 equal bins, each covering a 36° range. To adapt the histogram to 3D, spherical coordinates were used. The spherical anglespace was divided into triangular tiles of identical shape and area, identically arranged in space[1]. The old and new tiling divisions of the spherical anglespace can be seen in the illustration below:



Old New

As in 2D SIFT, here too, interpolation is conducted between the histogram bins to mitigate the effects of quantization. Specifically, at the orientation selected, interpolation between the vertices of the triangular tile, as presented in the following illustration:



3. **<u>Keypoint descriptor:</u>** Descriptor selection, too, is similar to regular SIFT, for the 3D case the descriptor is derived by taking a 3D region surrounding the keypoint and subdividing it into 4x4x4 sub-regions – for each of which a 12 vertice histogram is constructed as described, giving us a features vector with $4^3*12=768$ dimensions.

   As in 2D SIFT, a Gaussian window with parameter σ is used to give features nearer the keypoint a higher weight than those further away. Next, tri-linear interpolation is performed between the eight connecting sub-region centers. Thus, the added value to the bin at vertex v_i is:

$$f(v_i) = w_{win} w_s \lambda_i ||\nabla I||$$

   where w_{win} is the window width, w_s is the weight assigned the interpolated sub-region, and λ_i is the barycentric coordinate of ∇I with respect to v_i.

[1] *This is in contradistinction with previous approaches which divided the space into equal angles along a single plane - forming tiles of different shape and area - creating an unnatural and inappropriate extrapolation of the 2-dimensional histogram.*

Finally, the descriptors are L2 normalized, truncated by a constant threshold δ, and normalized again.

Let's dive in and look at the capabilities of the tool itself, you need to get SIFT3D from the website, we highly recommend you download the binary version:

SIFT3D consists of 2 main functions:

1. kpSift3D - Extract keypoints and descriptors from a single image.
2. regSift3D - Extract matches and a geometric transformation from two images.

Now let's look at two code snippets for these two functions:

```
% Load the image
[im, units] = imRead3D('data/1.nii.gz');

% Detect keypoints
keys = detectSift3D(im, 'units', units);

% Extract descriptors
[desc, coords] = extractSift3D(keys);
```

In this first, short code we'll read the image, call the **detectSift3D** function to get the keypoints, and the **extractSift3D** function to get their descriptors.

In our second example we'll show you the use of the features of SIFT3D to register two images. We have full MRI scans and cropped versions of those scans, and we are seeking to precisely locate the cropped MRIs' original position within the full MRI. Obviously, this can be done manually but it would be a time-consuming and tedious task. We will show you how to easily perform this task using a simple script with SIFT3D. In a nutshell, keypoints are extracted from the two images to be matched, and are then matched using the RANSAC algorithm.

```
% Load the images
[src, srcUnits] = imRead3D('data/1.nii.gz');
[ref, refUnits] = imRead3D('data/2.nii.gz');

% Register
[A, matchSrc, matchRef] = registerSift3D(src, ref, 'srcUnits', ...
    srcUnits, 'refUnits', refUnits);

% show the results (one keypoint)
pIdx = 5;

sx = matchSrc(pIdx , 1); sy = matchSrc(pIdx , 2); sz = matchSrc(pIdx , 3);
rx = matchRef(pIdx , 1); ry = matchRef(pIdx , 2); rz = matchRef(pIdx , 3);
sl1 = squeeze(  src(sx, :, :)) ; rl1 = squeeze(  ref(rx, :, :)) ;

subplot(2,1,1);
    imagesc(sl1)
    colormap(gray)
    hold on
    plot(sz , sy, 'R*')
subplot(2,1,2)
    imagesc(rl1);
    colormap(gray)
    hold on
    plot(rz , ry, 'R*')
```

Let's look at the script's code: the first 2 lines read the MRI scans, where the src is the cropped image and the ref is the full image.

Next, we call the registerSift3D function, which gets the following parameters, the ref and src image, each with its units, as received from the imRead3D function.

The function can have the following additional parameters:

- nnThresh - the matching threshold, in the interval (0, 1); default: 0.8.
- errThresh - the RANSAC inlier threshold, in the interval (0, inf); default: 5.0.
- numIter - the number of RANSAC iterations; default: 500.
- resample - if true, resamples src and ref to have the same resolution; default: false.

In the example below, we see the cropped image above, and the full image below. In red, we see one matched keypoint. You can see that the identified keypoint is located at corresponding locations in both images (the same location from the point of view of salient features). Also note, that the keypoint is located at an edge, making its features (the histogram representing the keypoint) stable and robust.

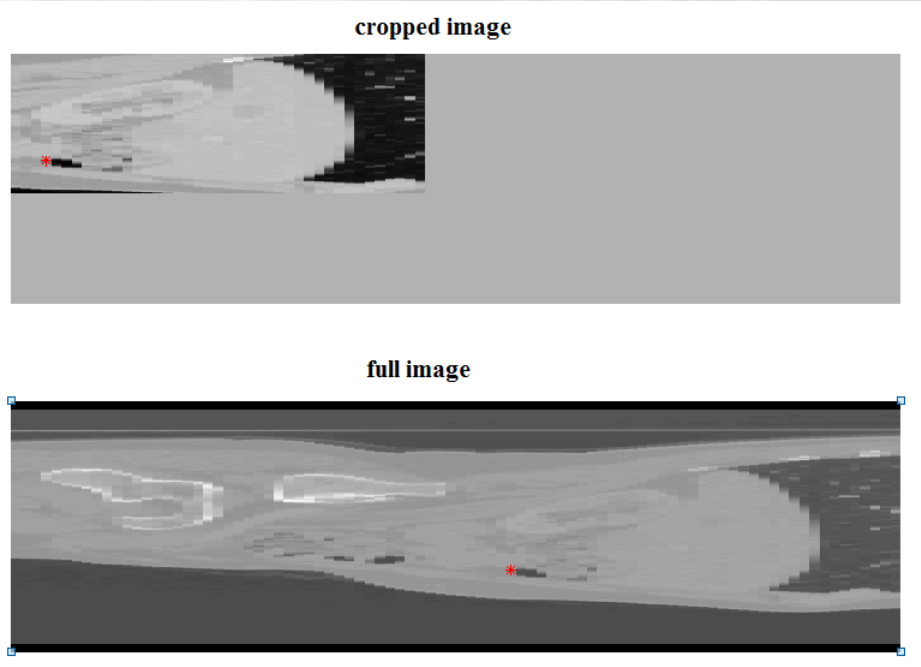


Image processing and linear algebra can perform file IO for a variety of medical imaging formats, including DICOM and NIFTI.

As a side bonus, the SIFT3D shipped with imutil, a library we found particularly useful for efficient saving and loading of NIFTI files into Matlab

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!

Identify overweight people from social media face photos

Let's start with one of those ideas that you never know if it's good or bad. It's about researchers from **MIT** and a **Qatar institute** who built an algorithm to identify overweight people based on their social media face photos; apparently, to alert those whose face presents patterns of obesity against risks of cardio-vascular diseases and diabetes. **Read...**

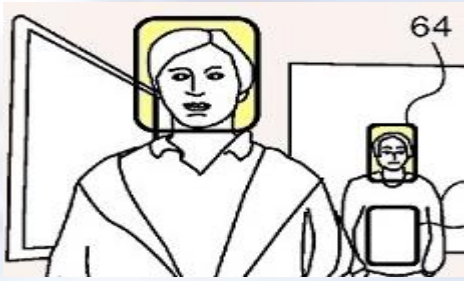
Salesforce Joins Computer Vision With Image Recognition Tool

Following last year's acquisition of MetaMind and as part of the launch of its new AI technology **Einstein**, **Salesforce** is also introducing **Einstein Vision**, a set of APIs that bring custom image recognition to CRM, allowing users to easily embed the power of image recognition in any app. **Read...**

Emotion Recognition in the Wild via CNN & Mapped Binary Patterns

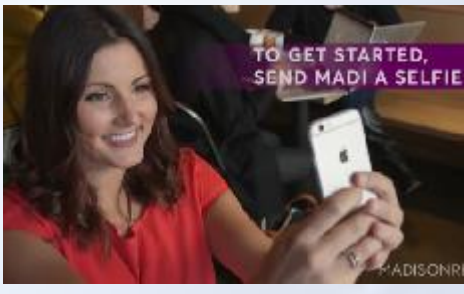
Flipping hamburgers is not the app that comes to mind in computer vision and deep learning. But this Pasadena, California restaurant decided to give a robot named **Flippy** the job of "**collaborative kitchen assistant**", in charge of grilling and flipping burger patties! **Read...** and **don't miss the video...**

Apple Granted Patent for Advanced Facial Detection



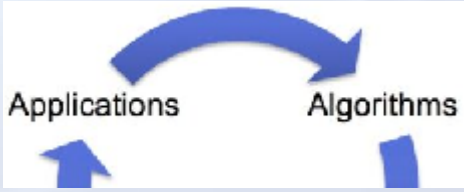
Apple is always granted many patents, but **Enhanced Face Detection using Depth Information** obviously caught our attention more easily than some magnetic buckle band for the Apple watch. This fine article gives some insight about Apple's **face recognition** system for desktops. **Read...**

Want Perfect Hair? Just Send Madison Reed Your Selfie



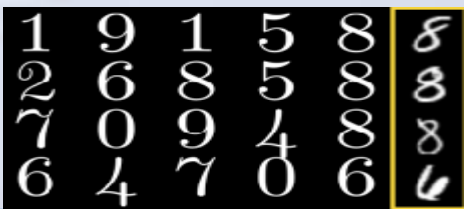
Madison Reed is an online shop for hair coloring products. Founder Amy Errett and her team designed a 12-question **quiz-based algorithm** to determine the perfect dye for each customer's hair. After women began sending their selfies, Errett turned to **computer vision** and developed a scalable app providing a quick and convenient response. **Read...**

Computer vision applications, both evolving and emerging



Nice report by **Embedded Vision Alliance**'s Editor-in-Chief, based on lectures and talks by some of their Summit's speakers indicating the trajectory of technologies which are becoming ubiquitous and "invisible". **Read...**

Stopping GAN Violence: Generative Unadversarial Networks



Let's conclude with something **funny**: **Read this paper...**

Tal Arbel - McGill University

We continue our series of interviews with **women in computer vision**. This new section, which we started with the CVPR Daily at CVPR 2016, hopes to help mitigate the severe gender imbalance in the computer vision community by getting to know better some **remarkable female scientists** and their career paths. This month, we interview **Tal Arbel**, who is currently Associate Professor in the **Department of Electrical and Computer Engineering, Centre for Intelligent Machines at McGill University** in Montreal, Canada.

Tal, where did you grow up?

I was born in and grew up in Montreal, Canada.

How did you discover that you had a passion for technology?

My father is an electrical engineer and he had his own business. Mine is a common path when you have a parent in engineering. He introduced me to programming at a very young age. It was an old TRS-80, and I really liked to program. He also let me solder. I was always interested in math and science, but through him I realized that I was very interested in engineering.

And that's why you started university in that field?

Yes, I studied Pure and Applied Science at **CEGEP**, the pre university college system we have here after high school. Then I studied electrical and computer engineering at **McGill University** for my

"We need more women in leadership roles"

Photo: Eva Blue

Bachelors. I didn't intend to go on to graduate school, but I got involved in some projects in robotics and computer vision and ended up staying for my Master's and PhD at McGill... so I did all of my degrees at McGill.

Were you ever tempted to choose another field?

No, I was very passionate about computer vision and interested in probabilistic methods and robotics, so I just continued along the project that I was excited in.

You are very much involved in everything. You chair conferences, teach many students, manage a lab and take on so many tasks. What makes you put so much on your shoulders?

I have many different interests. My main field of work is computer vision, but I became exposed to medical image analysis when I did my postdoctoral fellowship at the **Montreal Neurological Institute**. I realized that there are amazing opportunities to apply what I learned in the domain of computer vision to a wide variety of interesting problems in medicine where real improvements to healthcare can be reached. So I started to work in medical image analysis in the context of neurology and neurosurgery, developing new methods and providing tools to help clinicians, contribute to drug development and so on.

Is this what brought you to MICCAI?

Yes, I began to publish in more of the **MICCAI** field. I continue to publish in both fields and my students are working in both areas: computer vision and medical image analysis. I have recently become much more involved in conference organization and outreach in areas promoting women in science and engineering because I feel that it's important and it's an initiative that I'm passionate about contributing to.

We can bear witness to that from the MICCAI Women Networking Lunch in Athens. What do you take with you from that meeting?

That was our first open forum. **Gozde Unal**, **Ipek Oguz**, **Parvin Mousavi** and I put together this lunch at MICCAI to try to bring together women in the field from all different levels of their careers, from students to professors. We also had some male colleagues present and you. We tried to open it up to have a discussion about various topics, obstacles that women might be experiencing in terms of promotion and networking and so on. We didn't have very much time at that particular lunch. We put together a "**Sub-committee on Women at MICCAI**" whose mandate is to get more female scientists interested in fields of relevance to the MICCAI community, to try to get some policies in place that will encourage more women to enter the field, and to try to achieve more equitable career opportunities.

I remember two opposing opinions about the actions needed to improve the position of women in our community: one recommending positive

"In an ideal world, we wouldn't need these meetings promoting women in the field"

discrimination and the other viewing any assistance or benefit as a disservice to the cause. How do you reconcile these two views?

It was our first meeting, and that is a very common initial reaction. In an ideal world, we wouldn't need these meetings promoting women in the field, and everybody would have the same opportunities, but the reality is that we really don't. I think that asking: why do we need this? is a typical first thought. However, I think that there have been several successful examples of how these networking events can be very helpful, particularly to students and to new faculty members. I was involved in the first **Women in Computer Vision** meeting as part of **CVPR**. There, in addition to workshops where women presented, there was a successful dinner where female faculty members were teamed up in tables over dinner with PhD students in a mentor/mentee relationship. Students had an opportunity to ask about career path, express any obstacles they might be feeling and have that kind of relationship established. I had positive feedback from female students on this event.

Snorkeling in Bora Bora, French Polynesia

I think that at MICCAI, it's time to expose some of the issues that are common to many fields in science, in which women are underrepresented, expose implicit biases that exist whether we recognize them or not, try and create an opportunity for them to talk to other women and to network, and to hear speakers on the topic. Our particular sub-committee on women is working to organize a few more networking events.

"Oh! If I knew that was a career possibility, I would have gone towards science and engineering!"

If you could magically solve one of the discrimination issues, which one would you pick?

I think that in fields of medical image analysis and computer vision, we need more women in leadership roles. We are making strides to increase the number of women on the MICCAI board, to increase the number of women organizers at conferences and so on... but I do feel that **more leadership role models** should be encouraged. There are many excellent researchers in the field, so it shouldn't be too hard to create gender balance at the leadership level. This will encourage women, show that they are welcome and that they are able to attain and achieve success on par with their male colleagues. Overall, I'm quite surprised at the low percentage of women in the field of medical imaging, because medicine generally tends to attract a higher percentage of women. I think we need to do more outreach to promote

the field as a possibility for female high school students thinking about their careers. When I describe what I do to young women, they always say the same thing which is: "*Oh! If I knew that was a career possibility, I would have gone towards science and engineering!*" Personally, I'm chairing a committee on promoting undergraduate careers in Engineering at McGill where we are working on these issues, but there is still more work to be done.

Can you share a few words about your current work?

Sure. I head the Probabilistic Vision Group and the Medical Imaging Lab in the **Center for Intelligent Machines**. My work focuses on developing probabilistic graphical models and machine learning techniques in computer vision and in medical image analysis. I have been focused in the areas of neurology and neurosurgery. An example of recent work is where my students and I have developed machine learning techniques to automatically segment lesions and tumors from brain images of patients. I have been working with a local company (NeuroRx Research), that develops software analysis tools for drug development for Multiple Sclerosis clinical trials. The methods we have developed have been integrated into their system, significantly improving accuracy, cost, and speed.

In addition, my team is part of a recently awarded a multi-million dollar grant focused on progressive Multiple Sclerosis, led by McGill and including other universities worldwide, such as **University College London** and **Harvard**. My part of the project will be to develop automatic machine learning tools to find biomarkers in the images to predict progression in patients with Multiple Sclerosis.

I work on a number of other projects, including image registration in time-sensitive domains, such as neurosurgery, where I collaborate with the **Montreal Neurological Institute**.

"Those are the most satisfying parts of my job without a doubt: working with my students and helping with their career"

In terms of conferences, I am an organizer for MICCAI 2017, where I am Satellite events Chair. This year, we'll have 30 workshops, 14 challenges, and 4 tutorials. We have made several changes to the conference. One of the things that we have worked on for challenges is to try to streamline the process. Challenges are very important for the field of medical imaging where people don't necessarily have access to data. Challenges permit us to compare algorithms. You put data, ground truth, and metrics out there, allowing people to compare their algorithms on equal footing. It's extremely important in the field of medical imaging and we want to make sure that our challenges are fair and permit more scrutiny.

Uluru, Australia

I see a pattern here: you are so willing to help young women succeed in the field; you are so willing to contribute to the good health of people by helping medical research and drug research; you are so willing to help students progress. Is giving to others such an important drive for you?

[*laughs*] Yes, that is important to me. Those are the most satisfying parts of my job without a doubt: working with my students and helping with their career; I find the area of medical imaging incredibly satisfying and the fact that I am actually working on things that will be used on real patients and real procedures is very important to me. There are many really interesting problems. As long as there are interesting problems, I'll still be motivated to work on them.

What is the main lesson that you learned from your students?

I choose my students very carefully. I have excellent students and I love working with them in groups. My students come to me with papers, with techniques and with ideas. I consider them much more my colleagues than my students. The kind of students that I choose tend to love working in teams and to be very enthusiastic.

What is the greatest satisfaction you had from your students?

One of my students worked on Multiple Sclerosis. She worked on Gadolinium-enhanced lesion segmentation. Her thesis was applicable to a wide variety of problems including heart pathology segmentation. She won best paper in one of the MICCAI challenges on that topic. Her PhD thesis ended up winning the top PhD thesis award in computer vision in Canada at CRV, the Canadian Conference on Computer and Robot Vision. I love to see students come in, be very unsure, and just watch them develop confidence as researchers, develop passion, and see how different they are 4-5 years later… how they own their project, how they present it with such confidence, and apply it to so many different things. That gives me great satisfaction.

What would you like to achieve in your teaching or research career that you have not yet achieved?

In terms of career goals, I plan to continue to develop mathematical models and to apply them to many open problems in medicine. I'd like to develop machine learning and probabilistic algorithms to automatically learn what, in the patient images, are contributing to progression of disease and which patients would be more attuned to taking a particular treatment, resulting in tools for personalized medicine. I really believe that applying computer vision and machine learning to medicine has enormous possibilities in terms of improving patient outcomes.

Tal (right) with Zahra Karim-Aghaloo, author of the award-winning PhD thesis about detection of Gad-enhancing multiple sclerosis lesions

Waikit Lau, Arthur Chan, Samson Timoner

Artificial Intelligence & Deep Learning (or **AIDL** as we use to call it) is no doubt the surging **Facebook** group in our community. They will certainly deny it, but the main reason for their adding now about 1,000 new community members every week (!) is the perfect moderation put in place by curators/founders **Waikit Lau** and **Arthur Chan**. We have prevailed over their well-known modesty and were able to obtain an exclusive interview for Computer Vision News. **Samson Timoner**, who (among many other things) is also owner of another brilliant Facebook group (Computer Vision), joined the discussion moments later. What they tell us is a delight to read…

"Hey, let's start a group on Facebook!"

Why did you start this Facebook group?

Waikit Lau: [*laughs*] Why? I have no idea why! I'll give you some context on how this started. Arthur and I worked together at my last startup. Arthur was my resident machine learning expert in my last company which was in semantics indexing of videos, speech recognition and NLP. So a year ago I had this thought about how in AI and deep learning there's a lot of activity. There are a lot of things that are changing. I was talking to Arthur about this, and we were thinking that there wasn't anywhere we can go to ask questions and get responses fast or get the latest developments. We said *"Hey, let's start a group on Facebook!"* You know, chances are it will fail, and we won't get more than 100 people, but that's fine. We did it, and Arthur probably has done a lot of the heavy lifting in curating the group. People always ask us this question: how come your group is thriving so much compared to other groups? I say, there is only one reason:

A fascinating conversation on Cyclops! Clockwise: Waikit Lau (top left), Arthur Chan (top right) and Samson Timoner (bottom left) interviewed by Ralph Anzarouth.

Arthur Chan! Arthur spends a lot of time curating, responding, and being thoughtful about it. Otherwise, we would just become like any other group where people just post whatever they read elsewhere and no one comments. It's not that interesting. It's not really a community… so that's how we got started.

Arthur Chan: At this point, whenever we talk about the AIDL group, he loves to exaggerate my role in it, but Waikit is every bit as important as I am for the group because I remember there were multiple decisions such as how do we govern the group. All these fascinating things such as the YouTube "Office Hour" channel and the AIDL newsletter are Waikit's ideas, but he's a humble man. That's why we call ourselves "*humble administrators*".

I know you are both very humble, so I will ask you to put your modesty aside for my next question. The group adds about 1,000 new users every week. How many are we now?

Arthur Chan: Close to 14,000.

Why do you think it is so successful?

Arthur Chan: I think there are two major reasons. The first one is really about the deep learning craze that is still going on. People we talk with in the last couple of years believe that the deep learning trend will still go on for a few more years. The most important part is that the movement of deep learning actually covers multiple fields rather than just one or two fields. Even in the traditional deep learning fields such as speech recognition, computer vision and translation, we are still talking about how they are improving at a very rapid pace. I think that's number one because it is where deep learning is really going. Of course, there's always this existing base of people who are fascinated by AI. We have both of these two keywords [in our name]. The second reason is the very special strategies we have in curating the group, as Waikit mentioned. You can think at this this way: usually, there are two extremes of groups on Facebook. The first one is groups which don't have any rules. Everyone knows that there will be a tons of grabs and reposts. Some people put their nails on that site and claim that it's theirs. The other extremes are when the administrators are supposed to read all the posts before posting, but then there's no discussion in those groups at all. Our way reflects how Waikit and I see these kind of discussion. We let people in first, and then we exercise certain common sense judgment to decide whether things should be in the group. After the group has developed for close to a year, we developed a guideline which is very clear on what should be in and what shouldn't. A lot of members understand that there are certain posts you want and certain you don't want. I think that is the healthiest way, and

Arthur Chan

that is why we are still growing in a relatively fast way.

Waikit Lau: Yes, I agree with everything Arthur said. The one thing we cannot discount with all of this non-deterministic endeavor is luck. I think luck plays a big part in that maybe we got to a critical mass fast enough at the right time with the right set of folks in a critical mass, so more begets more.

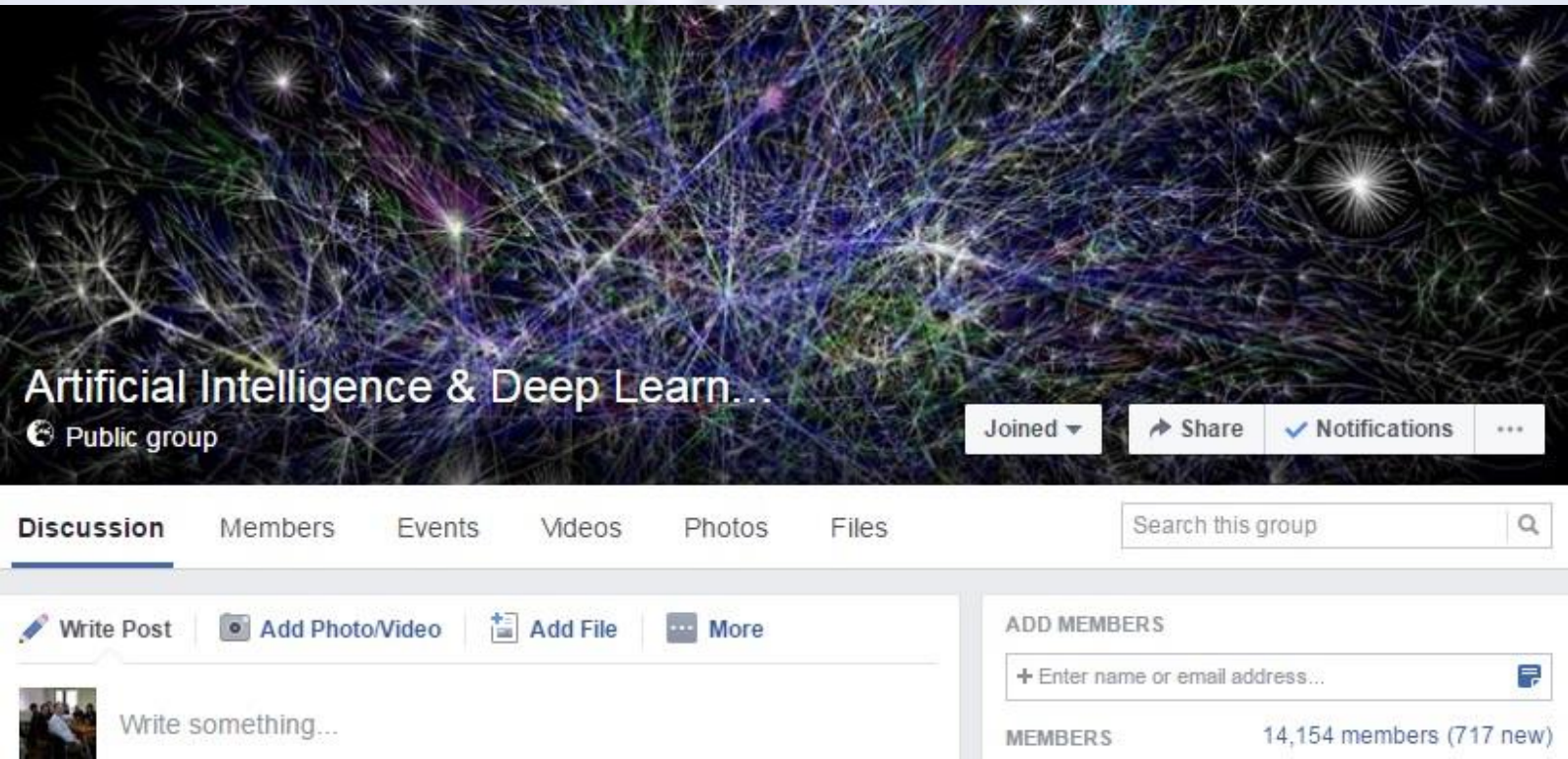
We often say "Wow!" when reading an impressive paper or learning about new techniques. Which was the latest paper or finding that made you say "Wow"?

Waikit Lau: There have actually been a bunch of them recently. I think in the past week, Kaiming He from AI research has a really interesting new version of instance detection and segmentation. The reason why I think it is so interesting is that you may think that instance detection, segmentation and characterization do not have a broad application, but that's what makes it really interesting. That's why people in machine vision are really interesting. Every little niche application can work well. What Kaiming has done is that he has taken what is essentially a niche segmentation, detection and classification problem and make it work really well. You can imagine the next time someone wants to do X classification or gesture classification, you can apply a similar thought process.

Arthur Chan: What I want to bring up is the timing of this [Kaiming's] research. I think the way I look at is a bit more technical: as every time people try to make a deep learning-based method trained end-to-end, the performance improved. That makes me feel that the Mask R-CNN paper is an important moment of the instance segmentation research.

This segmentation example is very new. But it happened to detection a couple of years ago. In the past, detection has been done in multiple stages and wasn't done end-to-end. As you know now, Faster R-CNN is one of the best, it is trained end-to-end, and you can adapt it to different classes.

You can see that trend in many fields. If you move this away from detection to segmentation, say speech recognition, people have been trying to train speech



recognizer end-to-end as well. Again, I think if it succeeds, it would be a major breakthrough. In any case, I think this kind of investigation of neural network architecture is fascinating for me. It's super AI nerdy and deep learning nerdy.

Today most of the deep learning researchers are employed by big companies like Google, Facebook, Amazon... Don't you think this might harm the diversity of future research in favor of certain industries?

Waikit Lau: I think that's actually a huge problem for AIs or its more generic machine learning developments. What is potentially the next industrial revolution powered by AI is now mostly controlled by a handful of large companies or organizations with research pointing at a very specific agenda. Google has done some very interesting stuff, just having a team work on cancer diagnostics and healthcare diagnostics, with larger budgets than the pharma and diagnostics companies; and startups cannot compete with all the Facebooks and the Googles of the world. It's a challenge. Hopefully this imbalance will correct itself over time, maybe as more and more venture capitalists and fund companies come in; they may not still pay as much as Google, but the thrill of building a startup and the thrill of solving a big idea would attract AI talent away from the Facebooks and the Googles into startups that are doing things that are maybe more worthwhile for humanity.

What kind of companies?

Waikit Lau: I'm thinking of healthcare. It's worthwhile, and those may not have immediate payoff now, but maybe 5 or 10 years down the road, all the way from discovery to diagnostics to whatever. There's so much impact, but that's not what Google and Facebook are doing. They are all about helping people to click more ads and things like that, which is great, but I think startups that work in the healthcare field and bio machine learning through health care will need to compete in a way that is non-monetary, like appeal to AI talent and say "*What we are doing is going to change the world. It's not just about the money.*"

Do you think young engineers can still be idealistic in their missions?

Waikit Lau: Yes, I think that as long as the company pays well enough, it's an incentive. After a certain level of pay, they may realize that making another $50,000 a year is not going to move me. I'd rather work on something worthwhile.

Which tutorial or book would you recommend to a new person joining the field?

Arthur Chan: That is the most frequently asked question in the forum. One of the things I wrote is an article called "Learning Deep Learning - my top 5 list": it is a kind of a starter list that you can follow on various things. Usually, it starts from a basic machine learning class such as Ng's. You can take other classes such as a statistics class. After one or two of them, you will feel comfortable with machine learning and can start deep learning. You can take Andrej Karpathy's class (CS231n). If you also feel comfortable with that, you can also take Geoff

Waikit Lau

Hinton's neural network class. That is a more difficult class in terms of concept. If you are smart, you can get a hang of it rather easily.

You have spoken about what others do. We would like to know what you guys do.

Arthur Chan: I have been an engineer on speech recognition for many years. I currently work for a small company called **Voci**, a speech recognition company. Waikit, Samson, and I are in the same area so we go to jam ideas all the time. The rest of the time, I administer the **AIDL Facebook group**.

What about Cyclops, the video-conferencing app we are using to talk right now?

Waikit Lau: Samson and I, with Arthur as an adviser, are working on Cyclops which is simple-to-use video conferencing. With video conferencing, you still have to download the package to use it. We said, as a start, let's make it super simple. With one click, you don't have to download: it just works in your browser.

This is what I did a few minutes ago.

Waikit Lau: Exactly. Was it simple?

Extremely simple.

Waikit Lau: Exactly! It ought to be that way. That's just a start. What if you could actually make video conferencing more productive? We think about what teams do in the same office. They go into the same room. They whiteboard in front of each other. They talk about ideas. It's very collaborative. People can't do that on video conferencing today using Google Hangout. The whiteboard is very washed out. You can't really see it.

We said "Why can't we solve this?" We said, on top of this video conferencing, let's build something that allows people to build applications and bots on top of it just like Slack. We want to open an API platform to allow people to build apps on top of it, so that people can be more productive. We started off by saying "*Let's build the apps ourselves first*". One of the apps we build, for example, is a whiteboarding app. Here, you can't really see too well, it's washed out. Now I turn on the whiteboarding app, and you can see it

more clearly. This is an enhancement, without which it is very blurry. The cool thing is that you can annotate.

Was it done by augmented reality?

Waikit Lau: Right - It has an augmented reality component in the sense that, in the real time, we are taking all of the pixels on the white board and we are basically enhancing the writing and the drawing. Then we can take this shot, and can send it to everyone.

We can toggle back and forth between white board mode and people. It also has an audio transcription mode. It is still in beta, and it doesn't work very well yet.

What were the challenges in building that?

Waikit Lau: Samson Timoner will answer that. He is a purist in computer vision and works with Arthur and myself at our company **Cyclops**.

Samson TImoner: First, as I'm sure you know, building a robust product that works across so many people's computers in so many countries is not an easy task. Not everything is standardized, and the things that are standardized still need to work. Your browser version is different than mine. Your computer is different from mine. On the white boarding part, we all have different cameras with different resolutions, different signal to noise ratios. You may have more or less processing power than I do. One of the real challenges is having an algorithm that runs smoothly on everybody's computer when you don't know how much power is going to be in their computer.

What is the tool that helped you the most solve this challenge?

Samson Timoner: That's a really great question. A lot of the standard web tools were great at profiling. It's what we used. A lot of those were simply built into Chrome. What it involves is getting feedback and literally walking into a store, trying it on a bunch of computers, taking a look at process managers, and seeing how we are doing. That's not an easy problem by any means.

"Learn how to trade stock!"

What's your advice for a young engineer entering the field?

Arthur Chan: My advice is, don't enter this field because it takes a huge amount of time to work on stuffs, wait for experiments... learn how to trade stock! [*everybody laughs*] Seriously, lots of people who want to work on data science feel like it doesn't take too much time. Mostly because it feels like you just work on math and think. The truth is you also have deadlines and some situations are hopeless. Sometimes you don't know how to improve the system more. Make sure that you prepare, and you learn everything. Don't stop learning, otherwise your skills will be stagnated very easily.

Samson Timoner: I'll add on the computer vision side that there's a big difference between making an algorithm work and shipping a product. I would advise anybody coming to this field to be part of a team and see the process... how an idea becomes an algorithm, how an algorithm becomes part of a module, then the module actually goes to shipping and it remains in a mode where people feel confident that it's working.

Waikit Lau: Go find a team that is A+. Frankly, it matters less about the company, because you learn from people. Traditionally, besides tier-1 companies like Google or Facebook, there are interesting, small companies doing interesting stuff. You also have to use due diligence and do some research. Plus, good chemistry is important.

Cyclops - Video conferencing with no download

Cyclops.io is the first video conferencing platform that uses computer vision. We used it to interview our guests (see pages 19-24) and we were able to have a very clean group discussion with four people, without any delays or no excessively metallic voices.

Cyclops has productivity apps on top, to replicate the in-the-same-room team dynamic. It works in your Chrome browser without any download or login. You can get up and running in seconds. Waikit Lau sent us **a link to click**, we clicked on it and we were **instantaneously in video conference** with him. We can assume that minutes later Arthur and Samson did the same to join the video conference.

It is great for short stand-ups as well as for hour-long meetings; you can also leave it as an always-on video/audio channel with your team. **Twitter**, **Houghton Mifflin**, **Sapient** and other companies are currently using the platform.

CYCLOPS

Apps and features:

(on top of the video conferencing platform)

- **Remote Whiteboarding:** if you tend to point your camera at a whiteboard during video conferences for team collaboration, Cyclops has a whiteboard mode that uses computer vision to dramatically enhance your writing so your remote teammates can read it. They overlaid annotation tools on it, so that everyone can mark up the whiteboard. You can post a screenshot to Slack or email it.
- **Audio Transcription:** Cyclops can audio transcribe in real-time, so no more taking notes in conference calls. It works best if you are using a headphone. You can post the transcript to Slack or email it.
- **Use it as an Always-On channel with team:** Unlike other solutions, Cyclops doesn't time out and reconnects automatically on Wi-Fi dis/reconnect. If you're on a different browser tab, it auto-pauses video but leaves audio running (no embarrassing moment or big brother feel).

The developers promise that more apps are coming...

Densely Annotated Video Object Segmentation (DAVIS)

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 have chosen to review the **Densely Annotated Video Object Segmentation" (DAVIS)**, organized around CVPR 2016 (held in Las Vegas) and 2017 (to be held this Summer in **Honolulu**). The website of the challenge, with all its related resources, is **here**. Read below what **Jordi Pont-Tuset**, a post-doctoral researcher in Prof. Luc Van Gool's **ETHZ vision group** and one of the organizers of DAVIS, tells us about this challenge (read also about Jordi's work **here** and **here**).

Background

Video object segmentation is about obtaining the binary masks of an object in all frames of a video, given its segmentation in the first frame. Applications of this can be object inpainting (removing one object and filling in the hole), interactive content visualization, etc.

In CVPR 2016, we presented the "Densely Annotated Video Object Segmentation" (DAVIS) dataset and benchmark, which consists in 50 full high-definition sequences annotated densely (every frame) with pixel-level unprecedented quality. It goes beyond previous datasets in video object segmentation in that it has significantly better annotation accuracy, it is much larger (thousands of annotated frames), its images are full high definition and it covers many real-world challenges (occlusions, fast motion, etc.).

Motivation

The scale and quality of DAVIS allowed the rapid deployment of deep learning techniques that boosted the performance of the state of the art significantly. Two recently-accepted CVPR 2017 papers (OSVOS and MaskTrack), for instance, reach performances around 80%. We quickly realized, therefore, that we needed to step up the level and so we created the **DAVIS Challenge on Video Object Segmentation 2017**.

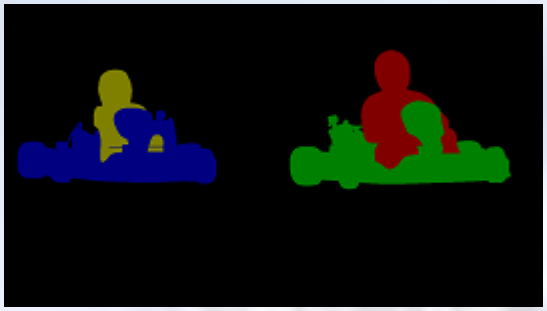
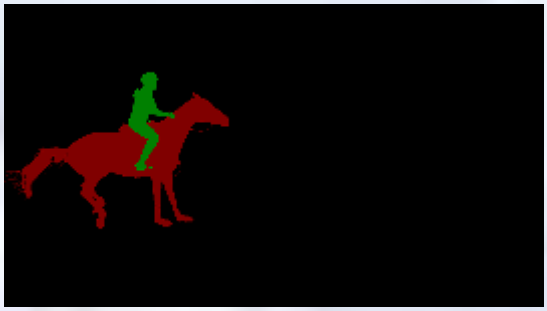
10.000 annotated frames, 500 annotated objects!

We have collected and annotated a new set of sequences with the following highlights:

- **150 sequences** (100 new) adding up to more than **10.000 annotated frames**.
- **Multiple objects per sequence** annotated, totalling around **500 annotated objects**.
- The majority of the sequences are **high-quality 4K UltraHD footage**.
- Same **pixel-level quality annotations** as the original DAVIS.

Jordi Pont-Tuset



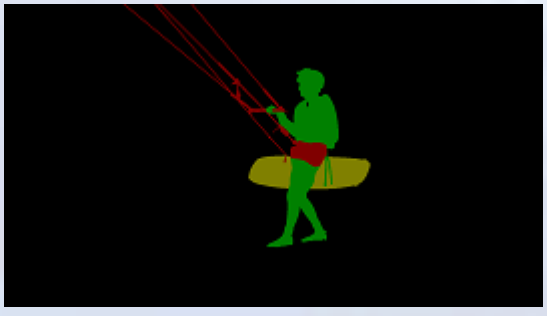
To promote research and competition in the field, we will host a challenge whose results will be presented at the CVPR 2017 workshop, that will work as follows:

- **April 1, 2017 - Open end**: Test-Dev 2017 phase, 30 new sequences, available 1st of April. Ground truth not publicly available, unlimited number of submissions.
- **June 19, 2017 to June 30, 2017**: Test-Challenge 2017 phase, 30 new sequences. Ground truth not publicly available, limited to 5 submissions in total.
- **Also on April 1**, we will make public 90 sequences and ground truth: the 50 original DAVIS 2016 sequences (reannotated with multiple objects) plus 40 new sequences.

The top entries will receive an **NVIDIA Titan X GPU** as a prize and will be able to present their work at the **CVPR 2017 workshop on July 26**.

We encourage all readers to consider participating and **good luck to everyone!**






Navigation System for Orthopedic Surgery

Every month, Computer Vision News reviews a successful project. Our main purpose is to show how diverse image processing applications can be and how the different techniques contribute to solving technical challenges and physical difficulties. This month we review **RSIP Vision**'s sophisticated **Navigation System for Orthopedic Surgery** based on advanced image processing algorithms designed to support surgeons in their task. RSIP Vision's engineers can assist you in countless application fields. **Contact our consultants now!**

Background

Not so long ago, surgery was simply the operation of a surgeon making cuts in the interested region, feeling with his/her fingers the bones being operated and fixing the mechanical supports which helped complete the task. There was no navigation system and as an alternative only sensorial touch and feel were available.

Navigation is a new concept and process: we use pre-op imaging of the patient, usually CT images, to position mechanical fixtures inside the patient's body without having to search for the specific location. CT image is used as **a map to navigate the body** of the patient.

What enabled **RSIP Vision** to achieve this major breakthrough is the integration of previously known techniques with our expertise in image analysis and our background in mathematics to perform calibration and registration.

Required Steps

Let's see what are the steps involved in this task; to do that, we shall first look at what **the final result** should be and at what is needed in advance (before the operation takes place) in order to achieve it. During the operation itself, we want the system to enable navigation across the CT of the patient.

We therefore need to know the **exact real-time location of every tool** (screwdrivers, surgeon's instruments and the like) in the real world, i.e. inside the body of the patient. This task is the **tracking of the tools**. We also need a registration of the tool with CT image. In addition, we want to locate the patient's body in the surgery room. Achieving all these goals requires several steps, like the following.

Example of CT segmentation in orthopedic surgery

Segmentation: CT scan is performed before the surgery. Our segmentation algorithm detects the vertebrae in the CT.

Calibration and Distortion correction: C-Arm, X-Ray portable camera is used to locate the patient in the surgery room.

The C-Arm scan distortion is corrected and the camera is calibrated.

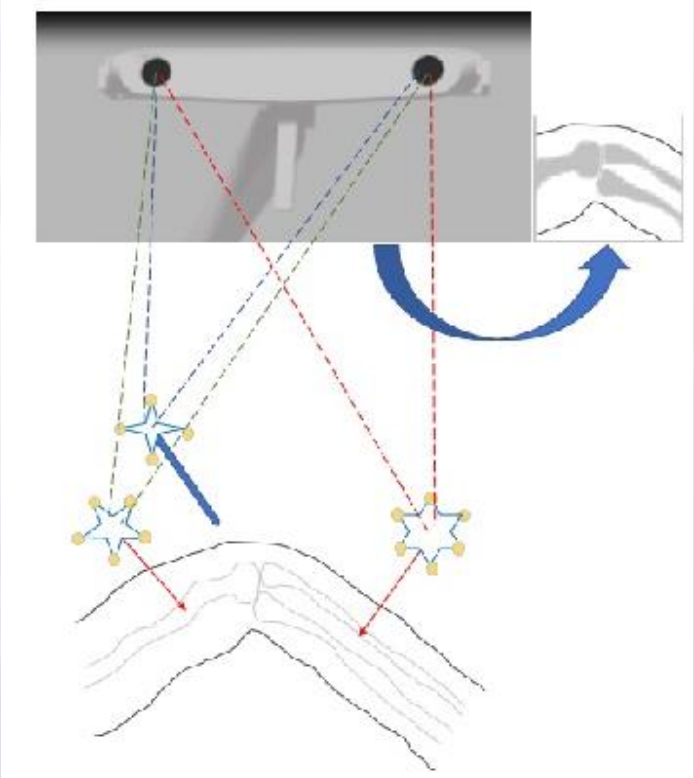
We superimpose the tool image on the CT itself, helping the surgeon navigate with the greatest accuracy

Surgery tool as seen on the CT scan

Registration: during the operation, we need to register the patient's 2D C-Arm scan with the 3D CT scan. Registration is done object wise, in the sense that actual bones are registered. This is why bones need to be segmented in advance. Additional registration needs to be done between the C-Arm and the tracking system, to know where the latter is currently located with respect to the camera. The registration between the tracking system and the 2D scan and after that the registration between the 2D scan and the 3D CT yield a transformation from the tracking system to the CT scan (The C-Arm is used as intermediate aid).

At this stage, **all is ready for surgery**: every movement of the tool will be tracked by the tracking system and its position converted into coordinates of the CT, so that we superimpose the tool image on the CT itself, **helping the surgeon navigate with the greatest accuracy**.

Each of the aforementioned steps has its own challenges. For example: segmentation requires to **analyze and locate each bone in all its details**, because knowing its precise position is key to achieve high accuracy. Registration too is a challenging task, because it requires **registration between bodies**, which is far from simple: every bone has its own shape (think at the vertebrae) and we have to make sure that this shape is perfectly matched. Also tools tracking needs to overcome **temporary occlusions** by the surgeon and other obstacles too. **RSIP Vision has all the experience needed to successfully solve all these challenges**. Find out how we do it in five articles about our **in-op navigation system**. You can read them at this link with many other projects by RSIP Vision in orthopedic surgery.



Steps in Computer Vision Software Projects

RSIP Vision's CEO **Ron Soferman** has launched a series of lectures to provide a robust yet simple overview of how to ensure that computer vision projects respect goals, budget and deadlines. This month we learn about the **Steps in computer vision software projects**.

This month we compare a general framework of project management to specific aspects of computer vision project management. We shall introduce the subject from a standard project management scheme and then show our recommended outline for computer Vision related projects.

A typical software-related project can be described as composed by the following 3 phases:

- Specifications - Project conception and initiation, project definition and planning
- Development - Project launch or execution, project performance and control
- Testing - Project close

Specification

During the first phase, project/product specifications must be received from the marketing department (or, if applicable, from sales) - [*for additional information, refer to SRS - Software requirements specification IEEE 830*]. Specifications should be converted into

R&D definitions, indicating challenges and risks. At this point, R&D prepares its own document, pointing out the alternative options for the development and the risks involved in each, including the time which will be potentially needed to handle those risks: the higher the risk, the longer its handling might prove to be [*SDD - Software design description IEEE 1016, SCM - Software configuration management IEEE 828, STD - Software test documentation IEEE 829*].

Once this is done, marketing and/or sales are presented with a solution, to obtain their approval to the proposed option, meaning that its objective is not far off from the desired result and that it matches the requirements [*SQA - Software quality assurance IEEE 730*]

When computer vision products are involved, it is important to **obtain as many sample images as it is possible to get**. These samples should span the whole range of the problem at hand and this should certified by the field engineers: the presented cases, stored in a dedicated data base, present immediate and long term datasets which need to be supported. Missing cases may degrade the long term's product performance.

The task of having all possible images for a computer vision project seems impossible. And indeed, we actually start with a limited dataset.

It is therefore important that the computer vision project manager understands the different complexities and the different directions that the examples database has to represent, as well as the quantities that need to be included in the database for each phenomenon.

Now, in the age of **machine learning** (as was the case in previous days, in the age of optimization and heuristics), algorithms parameters are numerous: a small set of images might limit the performance of the algorithm due to the narrow scope it represents. This is simply because **deep learning** is replacing the complex logical rules (coded in software) with automatic learning of the features in the image set (hence, deep learning). The larger and broader the image set, the better will be the image classification obtained.

At every milestone, specified objectives must be met following the predefined schedule

Development

Once the solution is agreed upon, it is time to set up the R&D plan. This plan includes phases and milestones, which clearly define the schedule and the expected deliverables through that schedule. At every milestone, specified objectives must be met following the predefined schedule. In case those weren't met, the development plan must be modified accordingly, assessing how this will affect the final product, in terms of capabilities, functionalities and delivery date. The project manager follows the development on a weekly basis and all the corresponding members should participate in the follow-up meetings and in the decisions, since these affect the R&D and (more importantly) the product itself [*SPM - Software project management IEEE 1058*].

This development phase and the follow-up performed by the project manager is more complicate in computer vision project management, owing to the very

wide solution options: for example, it is possible to add algorithm work to solve marginal cases. Project managers encounter a trade-off: they are forced to take into account ideas coming from the algorithm developers and at the same respect the product specifications designed in advance.

Each engineer might ask for more time to complete his task bringing new and ground-breaking ideas, so that the project manager is called to display the soundest professionalism in understanding the different options. At the same time he/she adopts a global view to comprehend the meaning of each and see the complete picture of all the development and of the resulting product. One example may be starting with a limited version of the product, by selecting only limited functionalities in some areas or opting for solving only a selected range of cases (albeit in a more rigorous and precise manner), in view of enlarging the scope of the product in a further release. This method enables to provide a high-confidence version, leaving the next level of challenges for the next version. In complex computer vision projects the strategy of starting with a simplified version built to demonstrate the method is recommended.

Make sure that the results are correlated to the desired project outcome

Among the many parameters which need to be taken into account, special attention should be given to those derived from the level of difficulty of input images, the level of certainty required from the algorithm and the level of desired results: making sure that the results are correlated to the desired project outcome.

The project manager might find quite challenging to deeply understand the computer vision algorithm developers on one side, and at the same time translate that in real world terms on the other side (Managers, Sales and Marketing). In other words, how a nuance in the algorithm or a set of parameters used to control it will potentially impact the final product. Without this understanding, the final product may be too complex to operate and potentially viable only for expert operators.

Note: in multidiscipline products, a final integration is performed here, at this stage. Multi-discipline project management will be discussed in one the next articles.

The real world imposes much broader cases that the product needs to handle

Testing

Once the final milestone is reached, the first complete version of the product is available and can be tested in the Alpha Site. It may be performed by running unit testing over a given examples database. Alpha is usually performed in-house, using field environment conditions: samples and operators mimic the field. Once it is completed successfully, a Beta Site testing is started. Beta site is performed in selected customer sites. The main customer benefit is that they can have an impact on the final products, nearly as if it was partly custom-tailored to their needs. In exchange, they offer their resources (time, expertise) to fine-tune the product and later test it in a real-world environment [*V&V - Software verification and validation IEEE 1012*].

The Testing phase has special characteristics which may be special to computer vision based products (and the like): **the real world imposes much broader cases that the product needs to handle**. You might have many bugs, which is natural when you expose a new computer vision system to a natural environment. Algorithm gaps may start to appear in the form of deficiencies resulting from the large data set including additional variations. The computer vision project manager should analyse the problematic cases carefully: evaluate their complexities and possible solutions. This information is used to set priority of handling or adopting alternative solution methods.

This process, consisting in both Alpha and Beta sites, might expose some parts of the system's limitations and bugs, which may even result in failure to meet the planned time-to-market. It is a very intensive period, during which the project manager ranks and prioritizes the different phenomena: doing this wisely enables to converge to the desired final product. Failure in doing so might cost precious time, preventing the meeting of the expected deadline. All participants should be involved in this step, to set the best solutions respecting the product specifications.

In the following articles, this section will discuss multi-discipline project management, common computer vision project methods and similar key topics.

YOLO9000: Better, Faster, Stronger Real-Time Object Detection

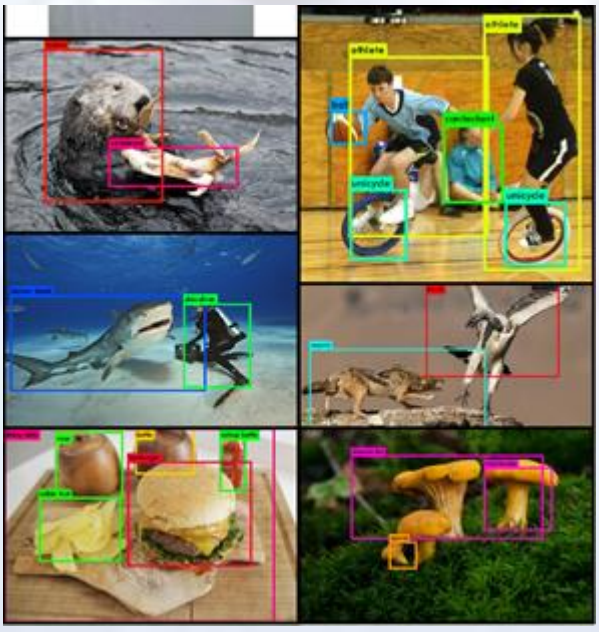
Every month, Computer Vision News reviews a research from our field. This month we have chosen to review **YOLO9000: Better, Faster, Stronger**, a paper proposing a state-of-the-art, **real-time object detection system** able to detect over 9000 object categories. We are indebted to the authors (**Joseph Redmon** and **Ali Farhadi**) for allowing us to use their images to illustrate this review. The paper is here.

Trained using a special new technique of training simultaneously on both the COCO and ImageNet datasets, which together include over 9000 categories

Prior detection methods are generally classifier-based systems: they apply the model to multiple locations and scales within a single image; they calculate the score for each individual region and elect high-scoring regions of that image as detections.

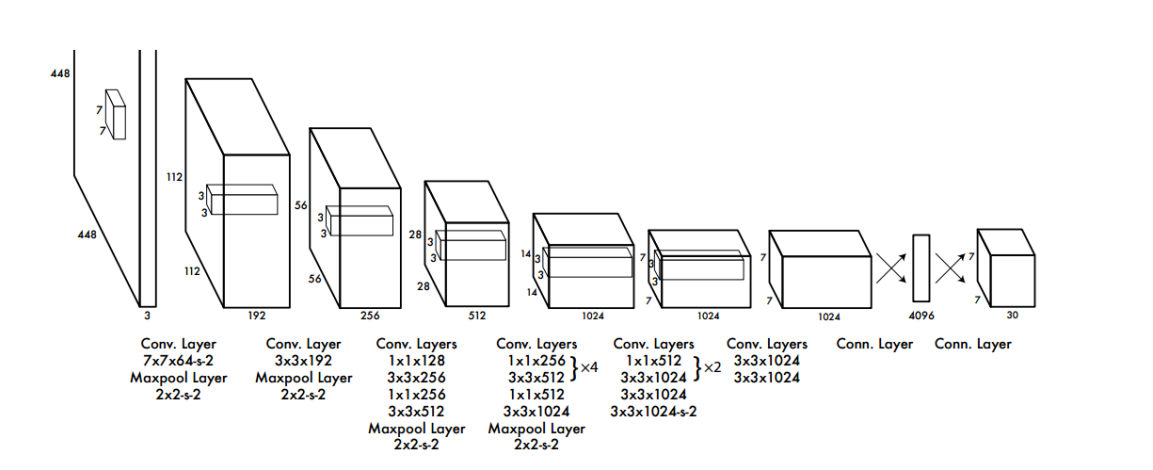
Our review will be focusing on two new versions produced by the YOLO team (YOLO stands for **You Only Look Once**), improved from what they presented at the CVPR conference in 2016 (which we will be referring to as YOLOv1): YOLOv2 and YOLO9000. YOLO9000 implements the authors' proposal to train a network simultaneously on object detection and classification; on the COCO detection dataset and the ImageNet classification dataset, which have over 9000 categories between them.

Here are examples of a few of the impressive results, from video(!):



Introduction:

The original version of YOLO from CVPR-2016 (YOLOv1 here on after), described in the illustration below, consisted of 24 convolutional layers, followed by 2 fully connected layers.



In YOLOv1 the image was divided into SxS cells of equal size, the cell that the center of an object occupies is responsible for detecting that object. Each cell is assigned to B (overlapping) bounding boxes; for each one, the cell has a score indicating the measure of confidence for that box, i.e. to what degree the algorithm is certain it includes a real object, and the accuracy it assigns to the category identification.

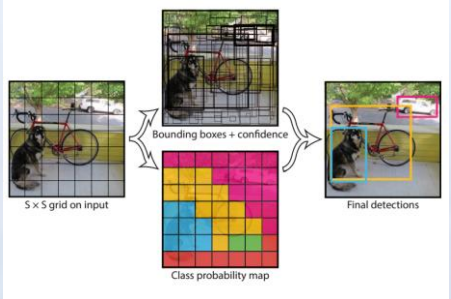
Each bounding box has 5 parameters: *x, y, h, w* and *conf. x* and *y* indicate its center, *h* and *w* indicate its height and width, and *conf* indicates the IOU between the (predicted) bounding box and any ground truth boxes in the image.

Moreover, each of the SxS cells estimates the probability of what object class is found in that cell; that is the function $Pr(Class_i|Object)$. Each cell of an image votes on only one class, regardless of the number of boxes it is assigned to. At test-time the system multiplies all the probabilities to arrive at a class-specific confidence score for each cell, according to the following formula:

$$Pr(Class_i|Object) * Pr\ (object)\ * IOU_{Pred}^{truth} = Pr\ (class_i)\ * IOU_{Pred}^{truth}$$

This score indicates both the probability of what specific class the object in the cell belongs to and to what degree the algorithm is confident of the accuracy of this classification.

This is summarized in the figure below: YOLOv1 identifies objects by means of regression. It divides the image into SxS cells, and for each estimates a bounding box B, and measure of confidence for each box, and a probability C for each class of objects. All of this data is saved in a tensor with the dimensions: SxSx(B*5+C)



Now, let's describe YOLOv2 and YOLO9000:

Method:

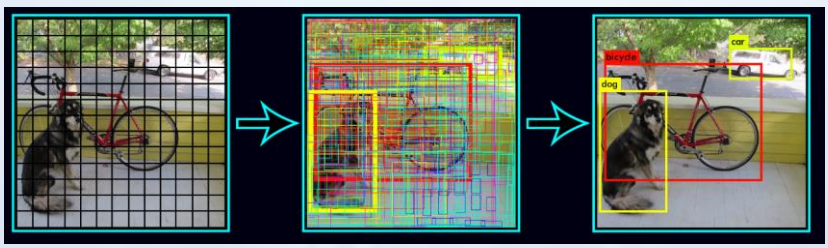
YOLOv2 is an improvement of YOLOv1; YOLO9000 has the same network architecture as YOLOv2, but was trained using a special new technique of training simultaneously on both the COCO and ImageNet datasets, which together include over 9000 categories.

YOLOv2

The basic network, called Darknet-19, was constructed based on YOLOv1 and similar research in the field. Like VGGNet it uses filters of size 3x3, and like NIN it uses 1x1 as well as global average pooling to make predictions. Its specific architecture is as follows:

Type	Filters	Size/Stride	Output
Convolutional	32	3×3	224×224
Maxpool		$2 \times 2/2$	112×112
Convolutional	64	3×3	112×112
Maxpool		$2 \times 2/2$	56×56
Convolutional	128	3×3	56×56
Convolutional	64	1×1	56×56
Convolutional	128	3×3	56×56
Maxpool		$2 \times 2/2$	28×28
Convolutional	256	3×3	28×28
Convolutional	128	1×1	28×28
Convolutional	256	3×3	28×28
Maxpool		$2 \times 2/2$	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Maxpool		$2 \times 2/2$	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	1000	1×1	7×7
Avgpool		Global	1000
Softmax			

YOLOv2 includes several key improvements over YOLOv1, which we will discuss now in detail:



The following table summarizes the percentile performance improvement per change.

- **Batch normalization** was added to the network and led to a significant improvement (faster convergence) at the training stage, and made other regularization methods (e.g. dropout) redundant. Normalization was added to all convolutional layers in YOLOv2 - leading to a 2% performance improvement.
- **High resolution classifier:** Current state of the art object detection networks rely on networks pre-trained on ImageNet. These networks usually run on input images smaller than 256x256. YOLOv2 is pre-trained on ImageNet data, at a resolution of 448x448, for 10 epochs. Then, as a second stage, it is trained to classify data into categories (already at the higher resolution). This process results in a 4% performance improvement.
- **Anchor boxes** were added to predict bounding boxes; and both fully connected layers of the network were removed.
- **Convolutional With Anchor Boxes:** YOLOv1 predicted the coordinates of the bounding box directly using the fully connected layers. YOLOv2, similarly to Faster RCNN, uses anchor boxes and offset directly on the convolutional layer, which authors admit slightly lowered mAP from 69.5 to 69.2; however, recall rose from 81% to 88%.
- **Dimension Cluster:** In the training process, when initial size of anchor boxes is set manually - if the manual selection is poor, though the training process compensates somewhat, mAP remains very low. Therefore, YOLOv2 uses the K-mean algorithm on the box, and centroid with a distance function to randomly select better initial anchor box size.

$$d(box, centroid) = 1 - IOU(box, centroid)$$

  where IOU is the intersection over union between boxes with centroid. Authors report k=5 gave a good tradeoff between recall and complexity of the model.
- **Direct location prediction:** To prevent a situation where every point of the image has a potential to be an anchor box regardless of location, even if it is at the very edge of the image. (With perfectly random initialization the model takes too long to stabilize.) YOLOv2's approach is to predict location coordinates relative to the grid of SxS cells. The network predicts 5 bounding boxes for each cell and 5 coordinated for each bounding box. The last two constraints introduced, the size of the box and its location - make the network's predictions more stable and easier to train, and led to nearly 5% performance improvement.
- **Fine-Grained Feature:** Adding passthrough layers that bring features from an earlier layer at 26 × 26 resolution. The passthrough layer combines higher resolution features with the low resolution features by stacking adjacent features into different channels similar to identity mappings in ResNet.
- **Multi-Scale Training:** YOLOv1 used input images of size 448x448, however, since YOLOv2 uses only convolutional and pooling layers it can resize on the fly. To achieve actual multi-scale training, the network would randomly select a new image size every 10 batches. The size-pool for selection was increments

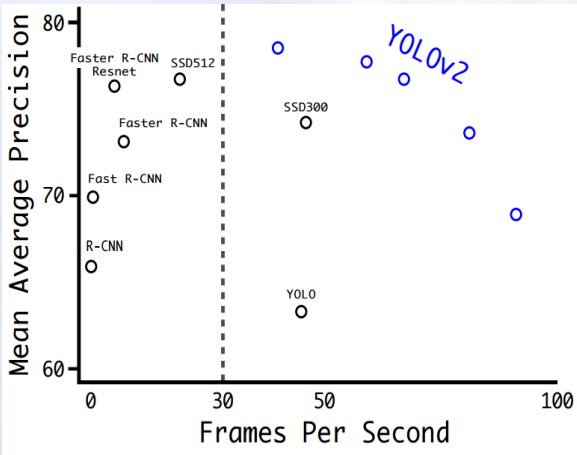
of 32 between 320x320 and 608x608, thus enabling the network to train on images of different sizes. At high resolution YOLOv2 is a state-of-the-art detector with 78.6 mAP on VOC 2007 while still operating above real-time speeds.

	YOLO							YOLOv2
batch norm?		✓	✓	✓	✓	✓	✓	✓	✓
hi-res classifier?			✓	✓	✓	✓	✓	✓	✓
convolutional?				✓	✓	✓	✓	✓	✓
anchor boxes?				✓	✓				
new network?					✓	✓	✓	✓	✓
dimension priors?						✓	✓	✓	✓
location prediction?						✓	✓	✓	✓
passthrough?							✓	✓	✓
multi-scale?								✓	✓
hi-res detector?									✓
VOC2007 mAP	63.4	65.8	69.5	69.2	69.6	74.4	75.4	76.8	**78.6**

As you can see, most of the network redesign decisions listed in the above table led to significant performance improvements (mAP).

Results:

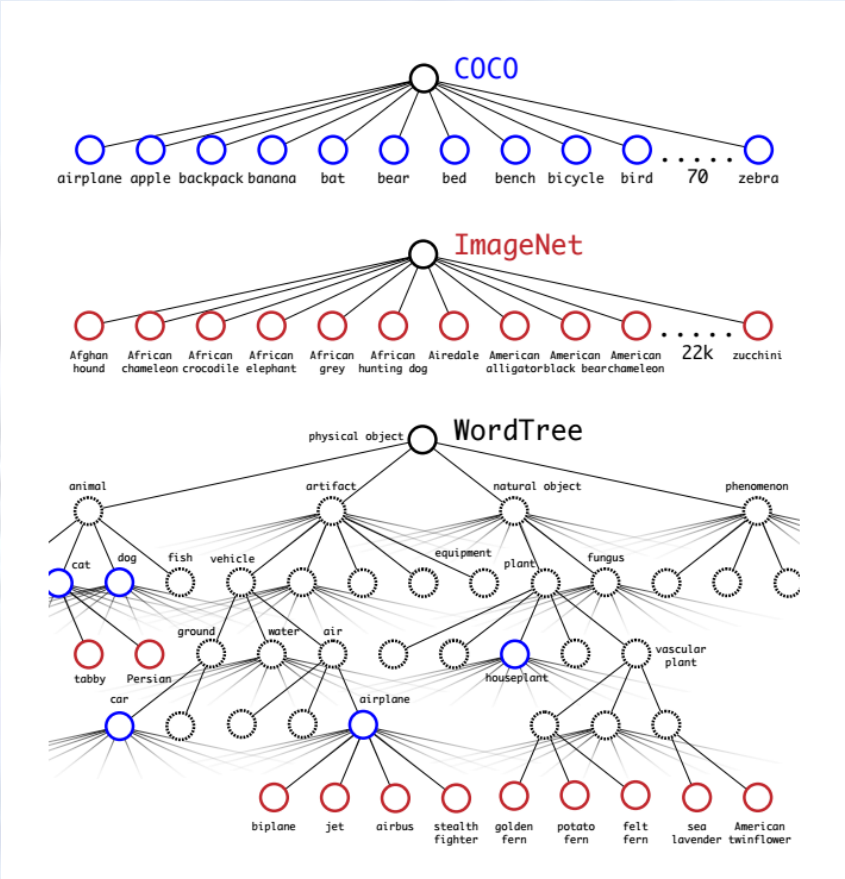
The results achieved by YOLOv2 on the VOC-2007 and VOC-2012 datasets are reproduced in the figure below. You can observe the tradeoff between image size, run-time and accuracy YOLOv2 offers, and in each run it achieves better speed or accuracy than all other algorithms (SSD, Fast R-CNN and Faster R-CNN).



Detection Frameworks	Train	mAP	FPS
Fast R-CNN [5]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[15]	2007+2012	73.2	7
Faster R-CNN ResNet[6]	2007+2012	76.4	5
YOLO [14]	2007+2012	63.4	45
SSD300 [11]	2007+2012	74.3	46
SSD500 [11]	2007+2012	76.8	19
YOLOv2 288 × 288	2007+2012	69.0	91
YOLOv2 352 × 352	2007+2012	73.7	81
YOLOv2 416 × 416	2007+2012	76.8	67
YOLOv2 480 × 480	2007+2012	77.8	59
YOLOv2 544 × 544	2007+2012	**78.6**	40

YOLO9000

YOLO9000 implements the authors' proposal to train a network simultaneously both for detecting objects and for classification. First, the COCO detection dataset and the ImageNet classification dataset needed to be combined: using the WordNet concept graph (which is diverse enough to cover most datasets' needs), a hierarchical tree of visual concepts was constructed, named WordTree. Below is an illustration of using WordTree to combine labels from COCO and ImageNet.



The authors used this combined dataset - WordTree - to train YOLO9000. The YOLOv2 network architecture just described was used. In training the network backpropagates loss as normal for image detection. For classification, however, only loss at or above the corresponding level of the label is back-propagated. For example, if the label is "dog" the network doesn't assign any error to predictions further down in the tree, "German Shepherd" versus "Golden Retriever", because it does not have that information

YOLO9000 was evaluated on the ImageNet detection task, the task shares only 44 object categories with COCO, meaning YOLO9000 has only seen classification data and not detection data for the majority of test images. YOLO9000 got 19.7 mAP overall, and an impressive 16.0 mAP on the 156 object classes not shared by COCO, for which it has never seen any labelled detection data. Note, it's simultaneously detecting for 9000 other object categories, running in real-time.

Conclusion and Demo:

The Authors introduce YOLOv2 and YOLO9000, two real-time detection and classification systems. YOLOv2 is state-of-the-art and faster than other detection systems across a wide range of datasets. YOLO9000 is a real-time convolutional network for detecting more than 9000 object categories by jointly optimizing detection and classification. WordTree - the combined dataset of COCO and ImageNet - was used to train YOLO9000. In addition, YOLO9000 is a strong step towards closing the dataset size gap between detection and classification.

Click for a demo to integrate.

Let's give it a quick try, just so you have a taste of how to use it: this demo can also be found at the official YOLO9000 website. It demonstrates how to use a pre-trained model for detection.

Install Darknet simply by typing:

```
git clone https://github.com/pjreddie/darknet
cd darknet
make
```

Download the pre-trained weight file:

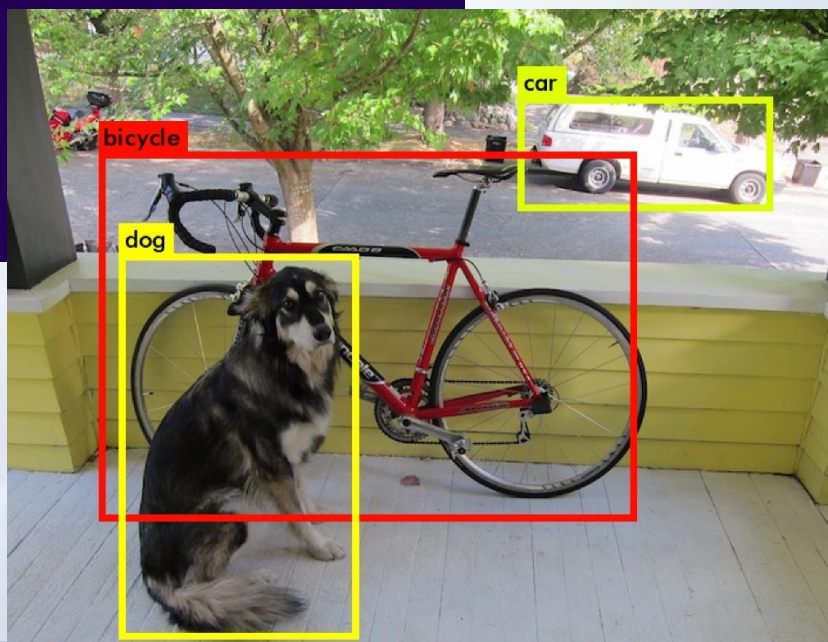
```
wget http://pjreddie.com/media/files/yolo.weights
```

Then run the detector!

```
./darknet detect cfg/yolo.cfg yolo.weights data/dog.jpg
```

You will see some output like this:

```
layer     filters    size              input                output
    0 conv     32  3 x 3 / 1   416 x 416 x   3   ->   416 x 416 x  32
    1 max          2 x 2 / 2   416 x 416 x  32   ->   208 x 208 x  32
    .......
   29 conv    425  1 x 1 / 1    13 x  13 x1024   ->    13 x  13 x 425
   30 detection
Loading weights from yolo.weights...Done!
data/dog.jpg: Predicted in 0.016287 seconds.
car: 54%
bicycle: 51%
dog: 56%
```



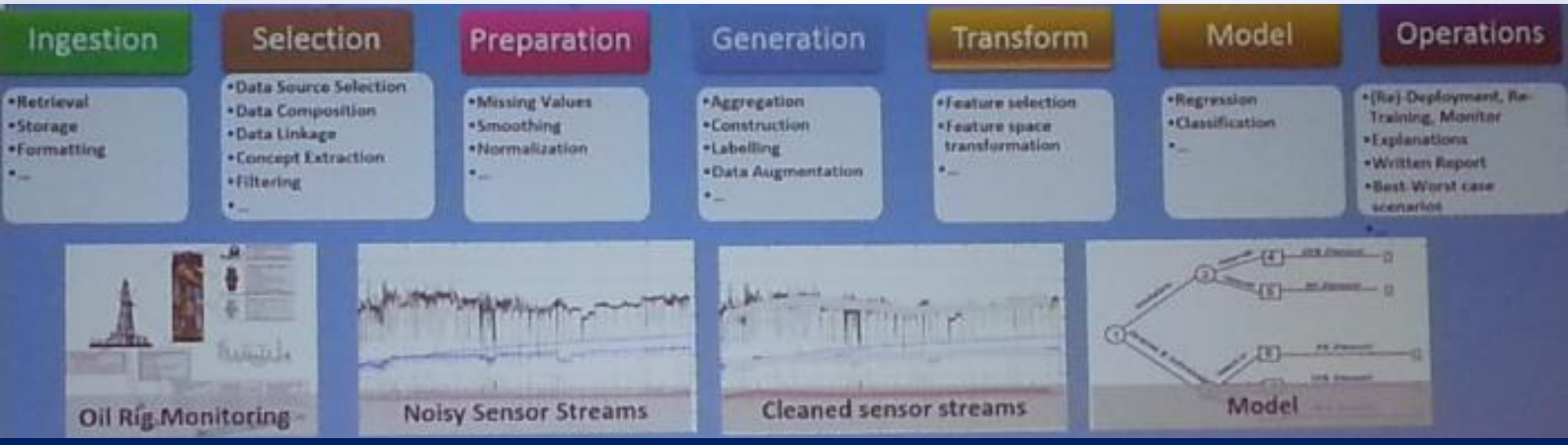
Click for more demos and full installation instructions.

Vision Monday Leadership Summit

The Vision Monday Leadership Summit took place in **New York** on **March 29**, with the title: **Supercharging Knowledge and Decision Making.** That was only a few days before the publication of **Computer Vision News.** However, thanks to our special correspondent **Lior G.**, we can show you a little bit of what went on there.

Session 4: A-Eye - How Artificial Intelligence is Transforming Eyecare

"There's a huge learning curve to applying AI to healthcare"
Pearse A. Keane, Moorfields Eye Hospital (UK)



From a slide by Chandra Narayanaswami, chief scientist and senior manager, IBM
"Analytic decision overload for Data Scientists" (content belongs to IBM Corp.)

CVVC - Intel

"How to test products that are based on deep learning and computer vision technology?"

Last year, **Intel** invited us to attend **CVVC2016**, their Computer Vision Validation Conference, in Haifa.

It was a great event and the technical program was exceptional. Among many others, we had the chance to follow the lectures of Prasad Modali, Danny Feldman and our CEO **Ron Soferman**, who gave an acclaimed speech on "**Computer Vision Validation - Key Learnings from Medical Devices and Other Industries**".

Intel is now announcing **CVVC2017** on September 11-12 at their premises in Haifa, Israel. The lectures will be in English and we recommend you give it a thought.

The conference program will be announced on July 2. Click on the image below to **register** (English info available at the bottom of the page). Read also about the **call for papers, tutorials and workshops**. Deadline for submission is on May 15.

Evostar 2017 - Bio-Inspired Computation
Amsterdam, Netherlands Apr. 19-21 Website and Registration

Machine Learning Prague 2017
Prague, Czech Republic April 21-23 Website and Registration

RE•WORK Deep Learning Summit in Singapore
Singapore April 27-28 Website and Registration

BIOMEDevice and Embedded Systems Conference
Boston MA, USA May 3-4 Website and Registration

CRV - Conference on Computer and Robot Vision
Edmonton AB, Canada May 16-19 Website and Registration

RE•WORK Deep Learning / Deep Learning in Healthcare
San Francisco CA, USA May 25-26 Website and Registration

WSCG Computer Graphics, Visualization and Computer Vision
Pilsen, Czech Republic May 29-Jun 2 Website and Registration

FG: IEEE Intl. Conf. on Automatic Face and Gesture Recognition
Washington DC, USA May 30-Jun 3 Website and Registration

AI Expo Europe Conference and Exhibition (with IoT Tech Expo)
Berlin, Germany Jun 1-2 Website and Registration

IEEE Intelligent Vehicles Symposium 2017
Redondo Beach CA, USA Jun 11-14 Website and Registration

SCIA 2017 Scandinavian Conference On Image Analysis
Tromsø, Norway Jun 12-14 Website and Registration

CARS 2017 Computer Assisted Radiology and Surgery
Barcelona, Spain Jun 20-24 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

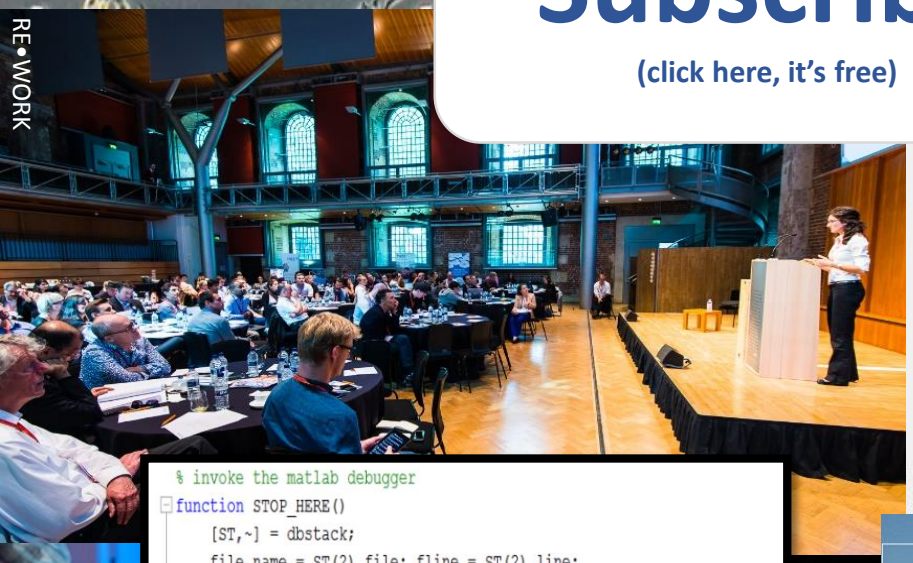
Computer Vision News

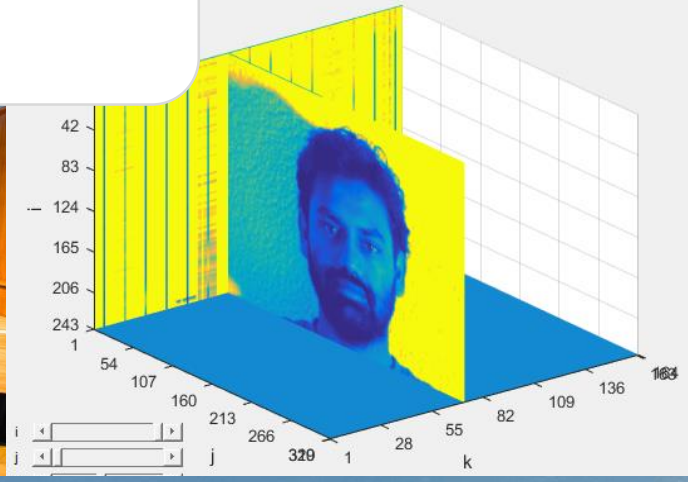
The Magazine Of The Algorithm Community

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

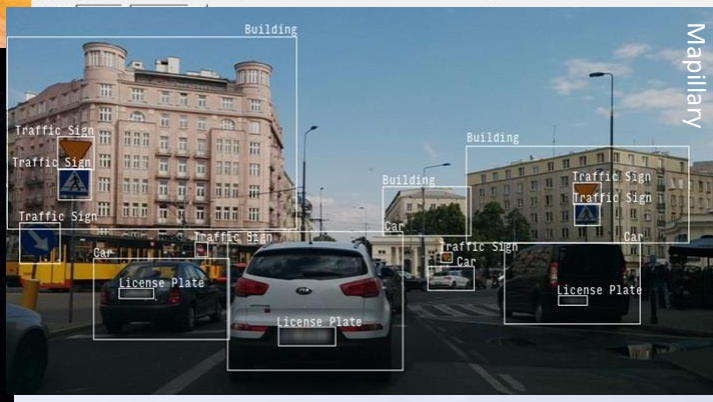
Subscribe

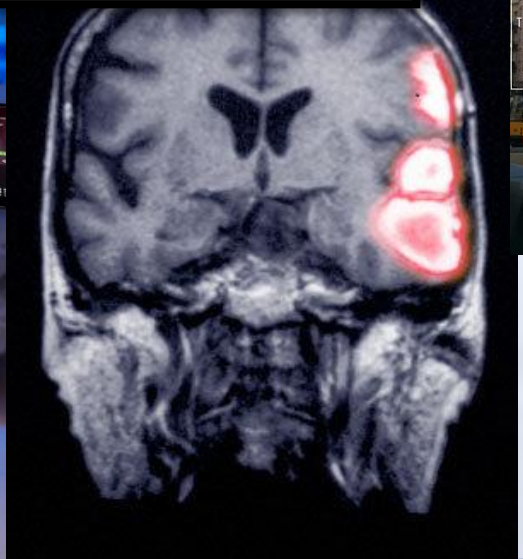
(click here, it's free)






```
% 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