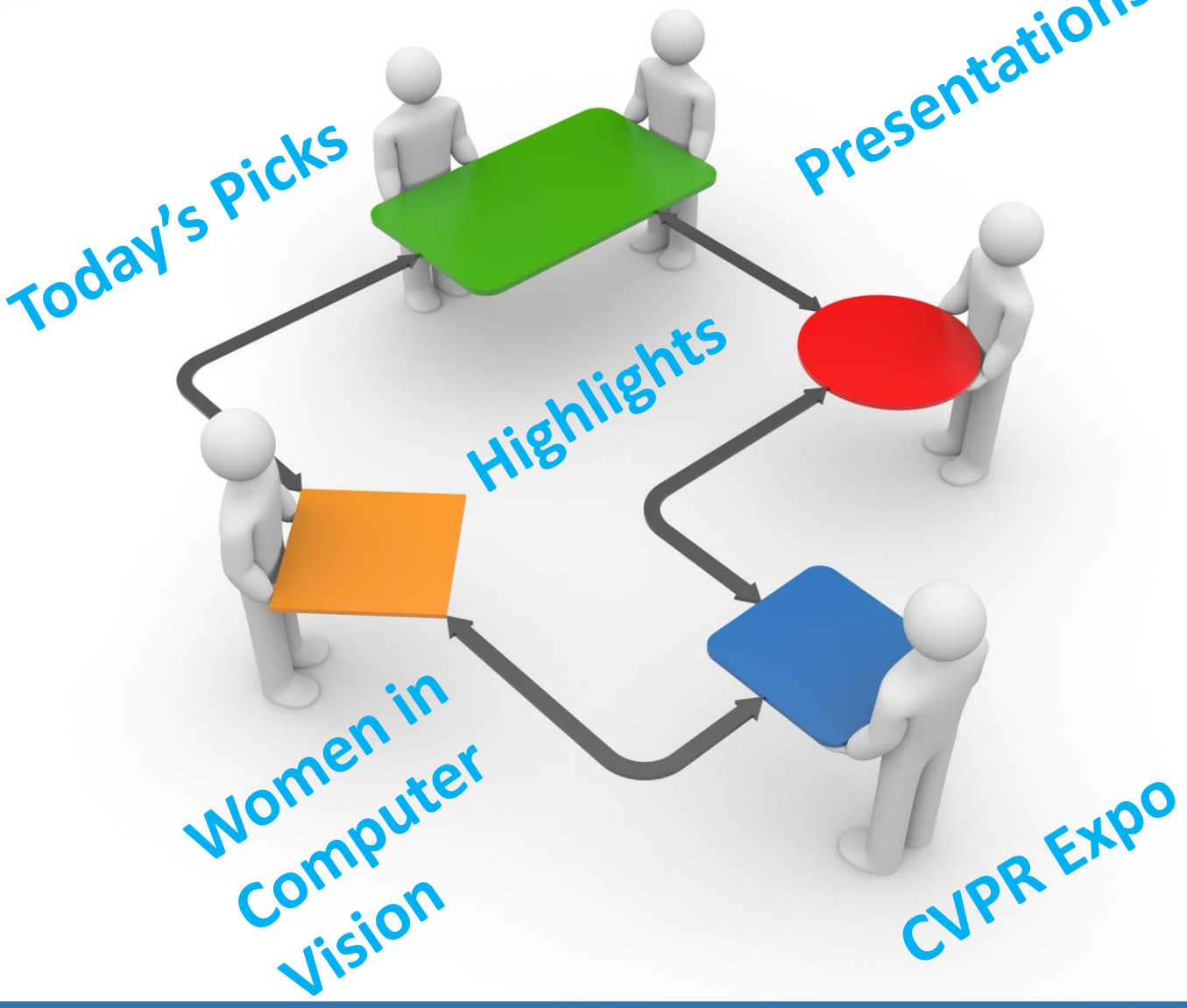
CVPR Daily

WEDNESDAY June 29



In cooperation with

Computer Vision News

The Magazine of The Algorithm community

A publication by RSIP VISION

For Today, Wednesday 29

David Pfeiffer is Principal Engineer at **Daimler AG**. He holds a Ph.D. in Computer Science from **Humboldt University of Berlin** and works in R&D for real-time computer vision and machine learning in the field of autonomous driving. David accepted to share his picks for today with **CVPR Daily**. Here are David's picks for Wednesday 29. Don't miss them!

Morning:

09:00AM to 10:30 AM
Semantic Segmentation
Oral session O3-1B-13:
Instance-Aware Semantic Segmentation via Multi-Task Network Cascades
Page 33 of the Pocket Guide and
presentation 13 on CVPR's website

"Multi-task network is really worth looking at: solving multiple problems looking at one only image"

Spotlight session S3-1B-20:
The Cityscapes Dataset for Semantic Urban Scene Understanding
Page 33 of the Pocket Guide and
presentation 20 on CVPR's website

"A great dataset to work with, exceeding everything there is in the field of pixel-wise, accurate semantic labels"

Afternoon:

1:45PM to 3:20PM
Video Understanding
Oral session O3-2A-1
Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled
Page 36 of the Pocket Guide and
presentation 1 on CVPR's website

"1 million hand images were not well labelled and the question is how to make sense of it. We do not work on hand gestures, but images are continuous, like video, and there might be good ideas that I might use in the automotive industry"

Video Analysis 2
Spotlight session S3-2A-11
Optical Flow With Semantic Segmentation and Localized Layers
Page 36 of the Pocket Guide and
presentation 11 on CVPR's website

"This is something we are working on right now…"

CVPR Daily
Editor: **Ralph Anzarouth**
Publisher: **RSIP Vision**

Dear Reader,

This is the third CVPR Daily and it is such a pleasure for me to discover every day how many readers we have, here at CVPR and outside.

In this issue of Wednesday, you will as usual discover the **recommended picks in today's program**. Have a look at page 2: we have good tips for you.

You will also read about presentations: namely, **Andrew Owens** presenting **Visually Indicated Sounds, Jing Wang** presenting the **Walk and Learn** model, **Muhammad Asad** presenting how to extract hand orientation from two-dimensional images and several others.

But yesterday's main highlight is certainly the excellent **Computational Photography and Faces** oral session, during which **Matthias Nießner** and his friends offered us a live video demo which created the buzz. Kudos to the Face2Face team for the nerve and guts. You will see in the next page a great image taken from that live demo. We hope you will like it as much as we did!

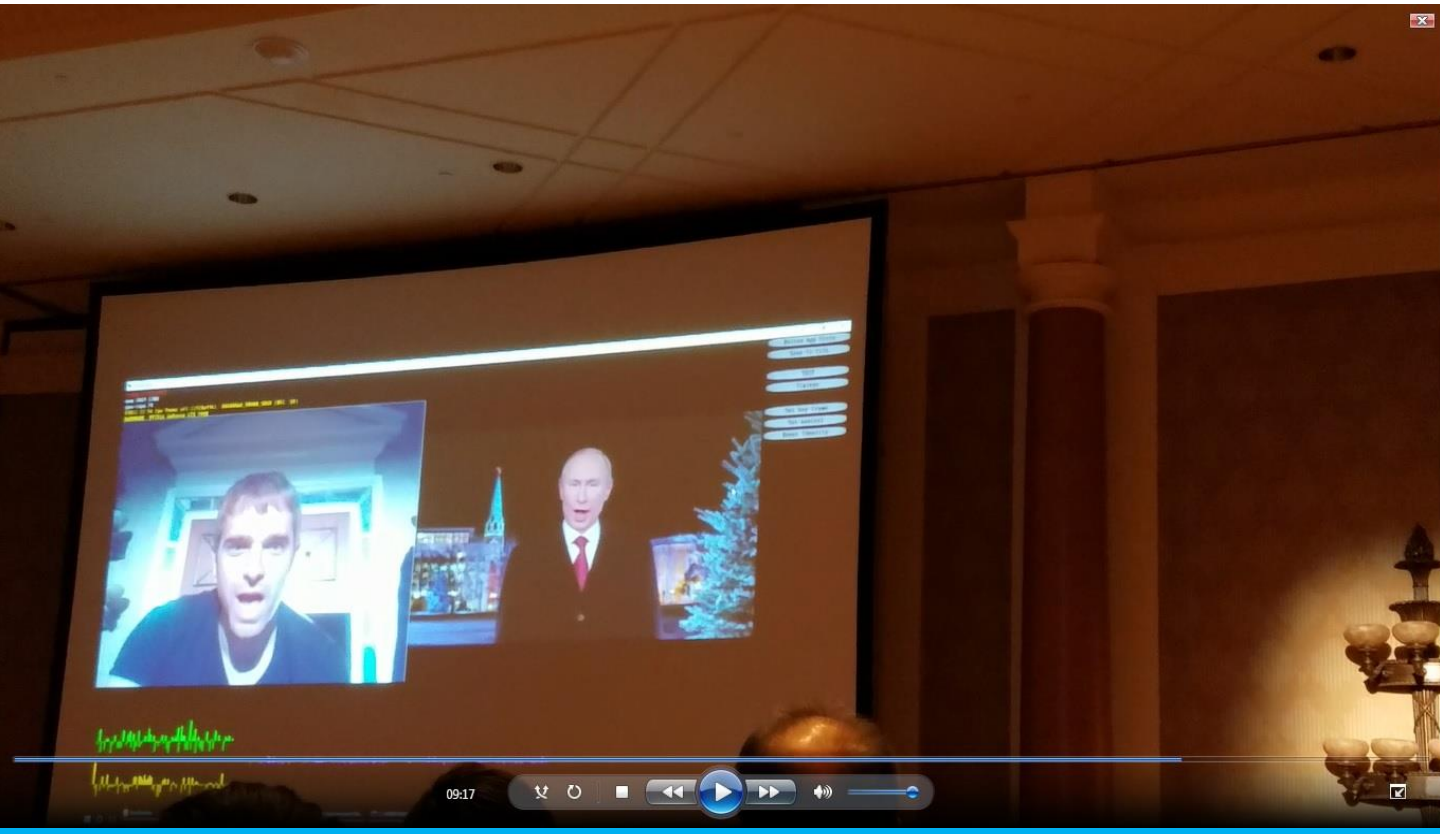
We wish successful presentations to all the presenters and a great day to all!

Enjoy the reading.

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

The greatest moment in yesterday's presentations was certainly the very brave **live video demo** that **Matthias Nießner and the Face2Face team** made at the **Computational Photography and Faces** orals (actually all presenters there were very impressive). The Face2Face research reenacted a monocular target video sequence (e.g. YouTube) based on the expressions of a source actor. The input to the method is two monocular video sequences: the first one features the target actor to be re-enacted and a second one features the source actor (e.g captured live with a commodity webcam). The output is a new video in which the facial expressions of the target actor are animated by the source actor in a photo-realistic fashion.

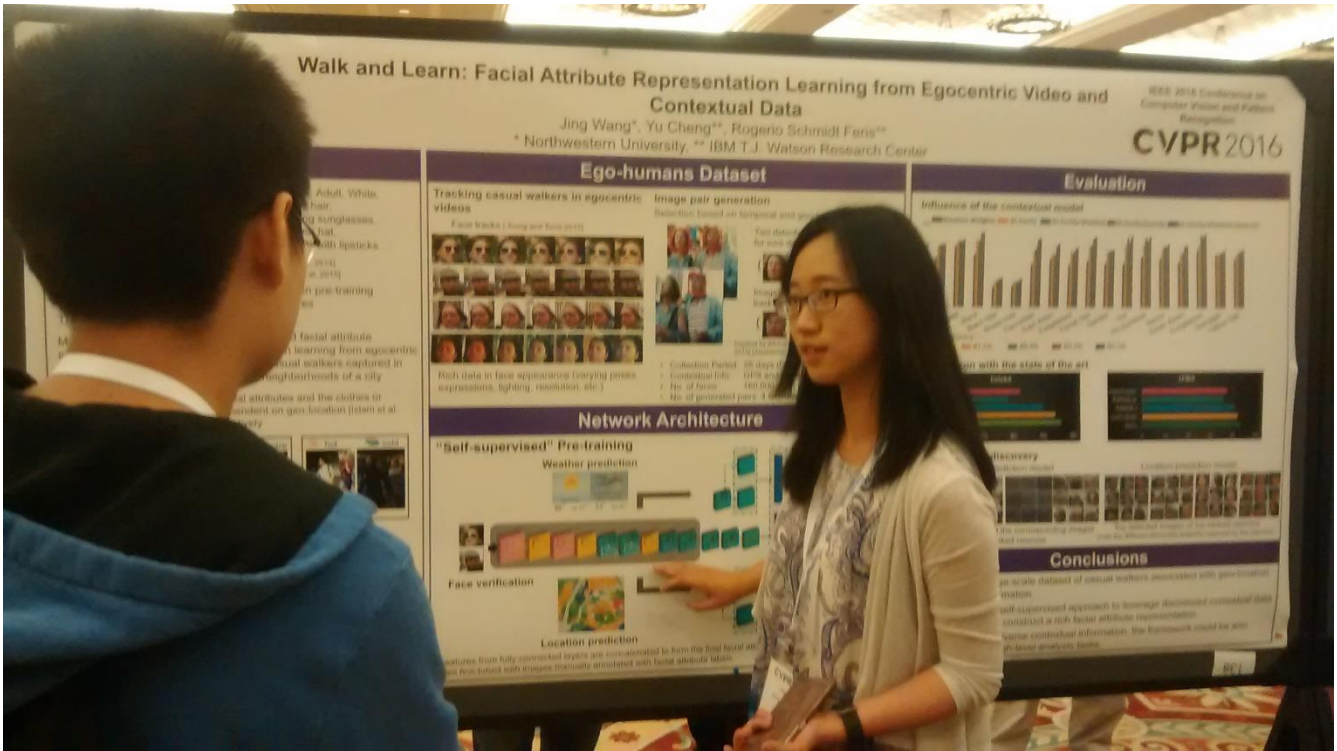
Yesterday's live demo source actor offered very funny face expressions to **Russian President Vladimir Putin**. The effect was so exhilarating that we were not even able to take a picture for our readers. I am therefore indebted to XING ZHANG (Mage Inc. /SUNY-Binghamton), for kindly allowing me to publish the one he took, reproduced below. You will not be surprised to learn that this work was already known to Computer Vision News, by which the Face2Face model was reviewed in the May issue. Read the full review here (pages 16 to 18).



How to offer one's own face expressions live to a very compliant V. Putin

Andrew Owens has presented **Visually Indicated Sounds**. In his words, the idea behind this work is to take a video where someone is hitting and scratching things with a drumstick. Then, to predict the plausible soundtrack to go along with it. The idea is that by predicting sound that's visually indicated in a video, or in other words, when you see the action that is producing the sound, then the algorithm that produces the sound has to implicitly learn about material properties where you're hitting. If you're hitting a carpet, it will make a very different sound than if you're hitting metal. The motivation behind this is to learn these interaction sounds in a way that might be similar to the way that humans learn. For example, people and children spend a lot of time interacting with objects and listening to the sounds they make. The team would like to take inspiration from the way they learn, to train computer vision systems that learn without explicit labelling.
The main technique used is **recurrent neural networks**: they take a silent video sequences' input. Then the recurrent neural network outputs the corresponding sound features for each frame.

Jing Wang presented the **Walk and Learn** model: this work constructed a **large dataset of casual walkers from egocentric videos** with weather and location information. Without requiring manual annotations, this project proposed a self-supervised pre-training model by leveraging discretized contextual information (geo-location and weather) as weak labels to learn features suitable to facial attributes. By learning with diverse contextual information, the framework could be also applied to other high-level analysis tasks.



Kuan Fang and Kevin Chen

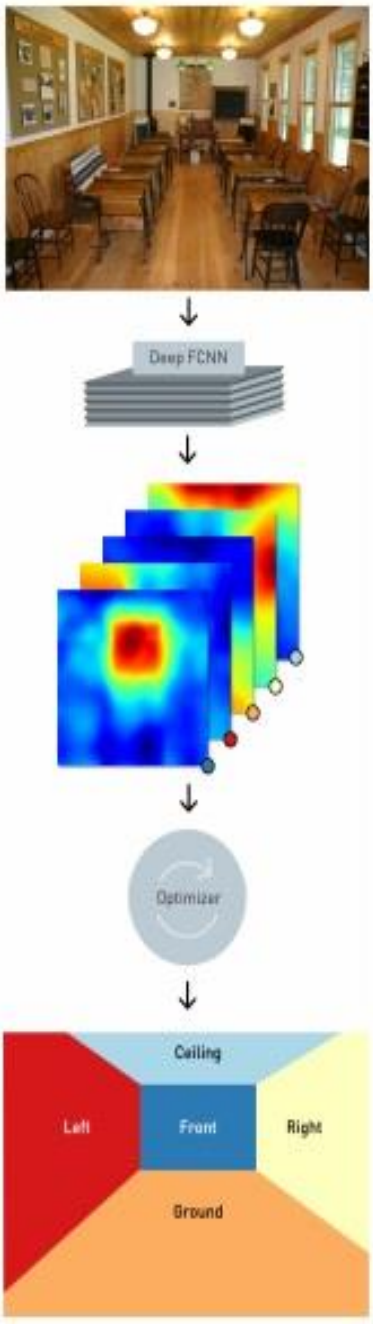


Figure 1: An overview of our layout estimation pipeline. Each heat map corresponds to one of the five layout labels shown in the final output. They are color coded correspondingly.

CVPR Daily: What is your work about?

Kevin Chen: The work is about taking a single monocular RGB image, and estimating the layout from just that one image. So the layout means you have for every pixel label of which wall in the image the pixel belongs. This is only for indoor scenes.

CVPR Daily: What is the novelty in this work?

Kevin Chen: The novelty consists in a completely different approach from previous methods, which try to generate these layout candidates then use a structured SVM to try to rank the layout candidates proposals. Instead of doing that, we treat this as a semantic segmentation problem.

We pass the input image to a **fully convolutional neural network (FCNN)** to generate an initial estimate of the layout. Then this layout doesn't conform to a standard room because a standard room has straight lines on the edges. So to alleviate this, we combine the FCNN with a novel optimizer that predicts the layout which conforms to a normal room layout.

CVPR Daily: What algorithms do you use in order to solve this problem?

Kevin: We used convolutional neural network along with an optimization that is reminiscent of coordinate ascent and uses logistic regression as well.

"We found ways to shrink a good neural network to give us a per-pixel probability of the semantic labels of the room "

CVPR Daily: What was particularly difficult in this work?

Kuan Fang: The traditional methods to tackle this problem are using vanishing point detectors. So they first have these geometric constraints and gather those vanishing points, lines, and edges. Then they get the 3D room layout.

We tackled this using fully-connected neural networks. The output of the neural networks doesn't have a geometric constraint. So we found ways to shrink a good neural network to give us a per-pixel probability of the semantic labels of the room and also have a corresponding refinement staff to observe geometric constraints to gather the final layout.

CVPR Daily: What is the next step in this work?

Kuan Fang: There are several steps. One thing is to estimate the layout from a single image, but we are considering using other techniques like **structure from motion** to estimate a more accurate layout. We are also trying to make it faster. So currently it will take 15-16 seconds for a single image, but we want to make it faster or even real-time... So that you can just turn on your camera, have a video of your room, and you generate a 3D layout in real-time That would be really cool and useful for a lot of things.

So potential application could be that you could use this 3D layout and 3D models for let's say IKEA or another furniture company, so that they can provide you with some 3D models. You can just augment reality using this 3D layout and 3D models. You don't want to buy furniture you don't like, then try it out in your room, and return it, but you can simulate it to get an **Augmented Reality effect.**

Muhammad Asad

CVPR Daily: Muhammad, where do you study?

Muhammad Asad: I study at City University London. I'm a PhD student in Visual Computing.

CVPR Daily: Where are you from originally?

Muhammad: I'm from Pakistan.

CVPR Daily: I understand that you have a presentation on Friday. What is your work about?

Muhammad: My work focuses on how we can extract hand orientation from two-dimensional images. This is relevant in augmented reality and virtual reality. I specifically focus on how you can build a model that can work out of the box and would enable people to interact and play with **augmented objects** as they do with real objects.

CVPR Daily: What is the novelty in the work that you are doing?

Muhammad: The novelty in my work comes from the multi-layered random forest structure which I came up with. It basically exploits the idea of having an ensemble of models, and how we can generate models which are expert at specific domains. However, combined, they still work to give an overall picture.

CVPR Daily: What are the practical applications of this work?

Muhammad: The application of my work includes **augmented and virtual reality**. I specifically look at how we can use hands to be able to interact with augmented and virtual reality worlds. It can also be used as a new type of interface to interact with different types of applications. It gives the user a natural way to interact with the digital world.

CVPR Daily: Can you please tell me what is the next step of your work?

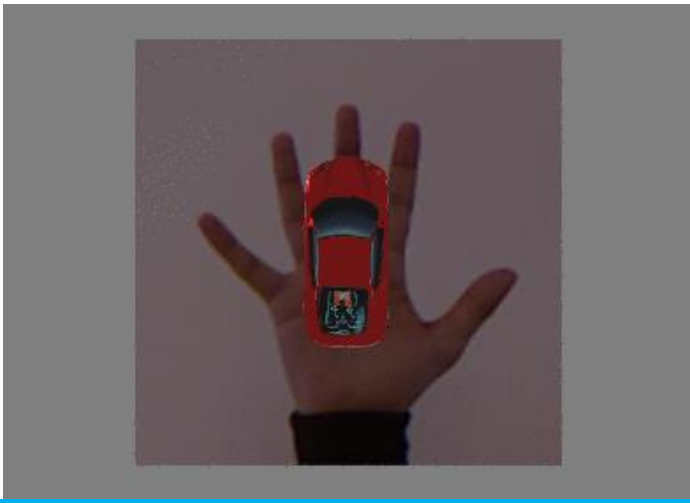
Muhammad: Currently, I have a manner that infers the orientation. In my future work, I'll look at how we can combine both pose and orientation of the hand and use that in **augmented and virtual reality.**

I'm also looking at building a temporal coherence. As the model evolves over time, how can we exploit the similarities and different movements of the hand and make the model more accurate?

CVPR Daily: Can you please tell me a funny story that happened during the development of this work?

Muhammad: My models are ambiguous. Specifically what happened is when I was testing them, if I move my hand up, it would show it going down. Basically the opposite happened from what I expected. It still is a problem in my model, but I'm looking to solve that in future work.

Augmented Reality on Egocentric Devices



Augmented Reality Using Hand Orientation

"Use hands to be able to interact with augmented and virtual reality worlds [...] It gives the user a natural way to interact with the digital world"

Anna Volokitin

CVPR Daily: Anna, where do you study?

Anna Volokitin: I am a first year PhD student at ETH Zurich, and I'm just starting a project on video summarization.

CVPR Daily: Can you tell us a few words about your work?

Anna: At this conference there is a paper called "**Video to Gif**" about the automatic extraction of gifs from videos. I'm working on continuation where we localize the videos. We also predict the gif boundaries.

CVPR Daily: Why did you choose to dedicate your career to computer vision technologies?

Anna: I think computer vision is close to cognitive science as well. It's so cool, and I am so fascinated by it.

CVPR Daily: What do you like about it?

Anna: I like that I can really see the progress in technology. It is moving so fast, and it's a really exciting time**.**

CVPR Daily: Is it more difficult for a woman to be a scientist and if yes, why?

Anna: Sometimes I second guess myself because there are no other women in our lab. In this way it's more difficult, but I think that in my case all of this criticism is only coming from myself, since everybody I have met has been super supportive.

CVPR Daily: What would you like to achieve in your career?

Anna: I would like to solve a real problem that people are having in this field. I am interested in human cognition and how it is related to computer vision. So if there is any insight that is between the two, then this is something that I would like to work on. I think that computer vision is going to go beyond human vision. I think first we have to get to human vision.

CVPR Daily: What will happen when computer vision will go beyond human vision? What will the world look like?

Anna: I don't know! It's going to be science fiction!

CVPR Daily: What would you like to do in that regard?

Anna: I think the most interesting thing will be to have insight into how to represent thoughts. If computers will have thoughts the way we do because they can also see and think about things.

CVPR Daily: Can this happen?

Anna: Why not?

CVPR Daily: Will you be the one who makes it happen?

Anna: It will be a team effort.

Khurram Soomro

CVPR Daily: Khurram, what is the work that you are presenting?

Khurram: I'm presenting a work about "Predicting the Where and What of Actors and Actions through Online Action Localization". The basic task is to predict what the action is and where it is happening in a video. Let's say we are streaming a video. We want to say what the action is and where it is happening by looking at each frame.

CVPR Daily: Can you tell me what is the novelty in this work?

Khurram: The novelty is that it's the online way of doing it. We are the first ones who can predict the action and localize it as we are viewing the video. Traditionally, the way that people do the action localization is in an offline manner. I give you a video, then you localize and detect it. However, with our work, we are able to predict what the action is and detect it while you are viewing the video.

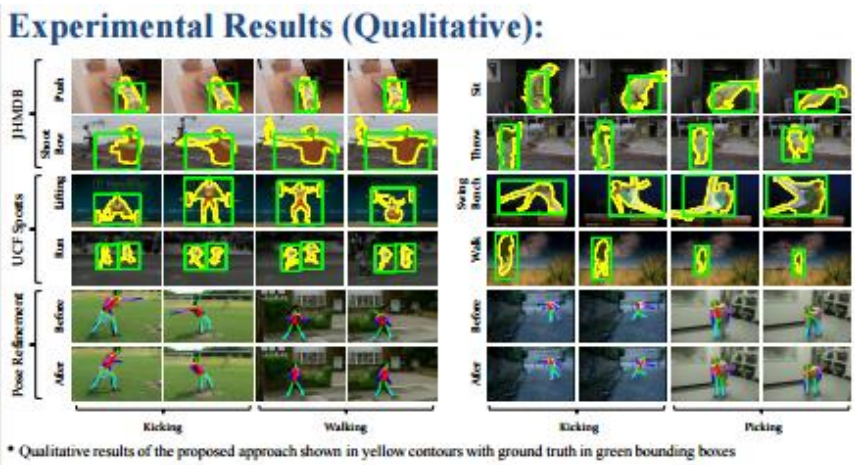
CVPR Daily: What is the main algorithm that you used in this work?

Khurram: The main algorithm is that we are using a combination of super pixels and poses together to learn a foreground likelihood model that distinguishes the foreground object from the background. We use super pixels within the pose bounding box to come up with a way to assign scores to each super pixel. Once we use the poses and super pixels together, we refine these poses.

CVPR Daily: Can you tell me what was particularly challenging in this work?

Khurram: The challenging part is that because we are going to predict the action, we have limited information as we are viewing. Let's say we are only seeing 10 frames of a 100 frame video. You are only basing your prediction on those 10 frames. That's why it becomes more challenging compared to offline manners which have the entire video and the entire motion visible. It becomes easier in that case.

CVPR Daily: What are the practical applications that this work can generate?



* Qualitative results of the proposed approach shown in yellow contours with ground truth in green bounding boxes

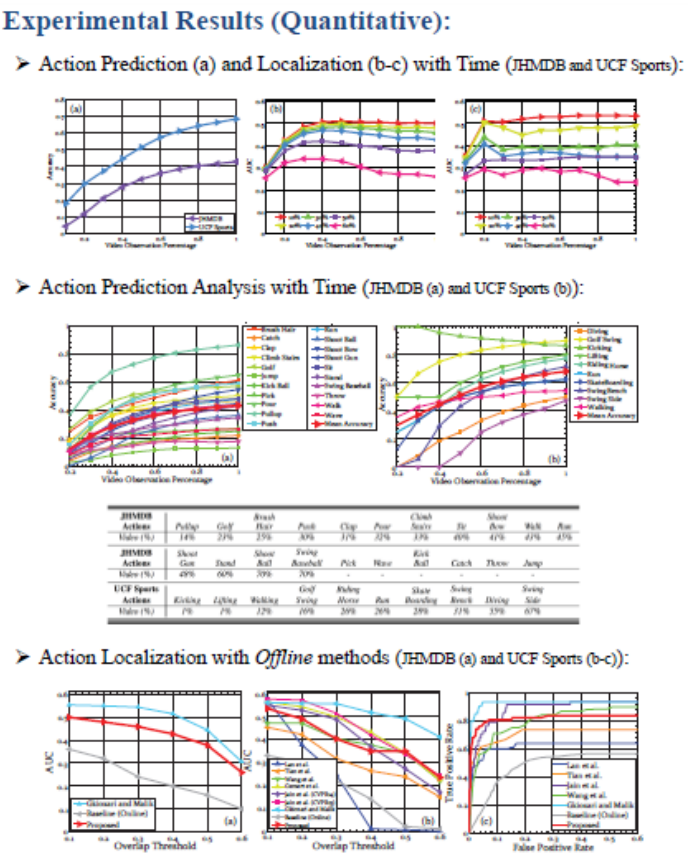
Khurram: The practical applications are huge. For example, this can be used in surveillance tasks. Now you have millions of CCTV cameras all over the world. However, there are very few methods which can detect in an online manner. Let's say I'm viewing footage from a CCTV camera, can I say what's going to happen as I'm viewing it? That's one aspect of it.

The second aspect is in human computer interaction. Let's say you are interacting with a computer in a Playstation or an Xbox. Now the action that is being performed in that interaction depends on what you do. This interaction needs to happen in a live manner. This can be used so you can estimate what the computer is doing, and the computer can know what you're doing.

CVPR Daily: What is the next step?

Khurram: The next step is to find a real-life application and make it work so that people can use it. It was quite interesting and challenging when I started to work on this because this is a very new problem. People are not doing it in an online manner which is very necessary for the application. This gives a huge area that people can start working on.
One interesting analysis that we can come up with using our method is that since you are predicting what the action is, we can say how much of the video you need to watch if you are looking for a certain action. If I'm looking for a certain action, let's say kicking a ball, how much of the video do I need to see to be able to predict what the action is? We do the sort of analysis that can help and distinguish which actions are more challenging than the other ones.

"since we are predicting what the action is, we can say how much of the video you need to watch if you are looking for a certain action"



Improve your vision with

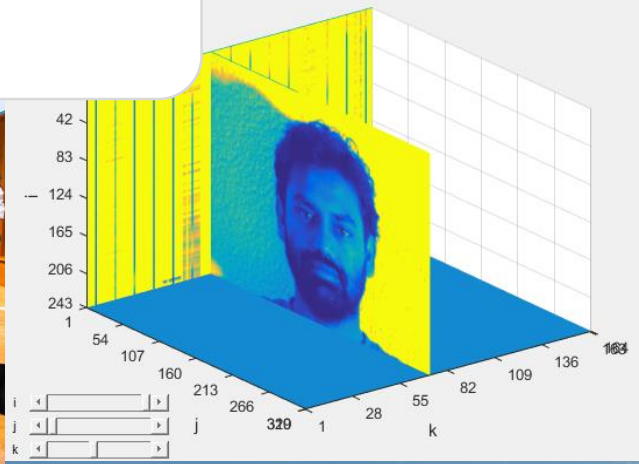
Computer Vision News

The Magazine Of The Algorithm Community

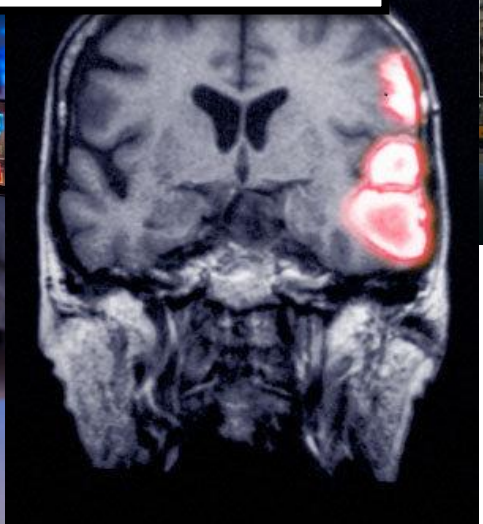
The only magazine covering all the fields of the computer vision and image processing industry

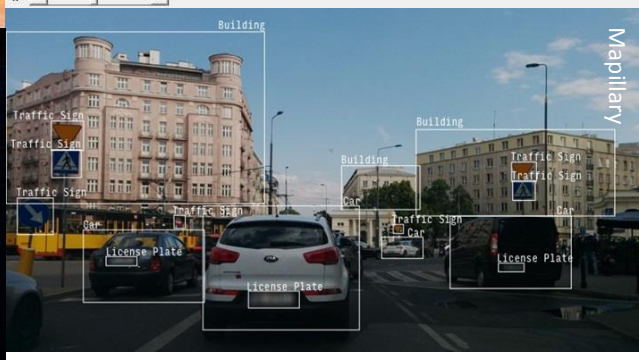
Subscribe

(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