SHIMON ULLMAN: Now for a different, entirely different type of issue that has more to do with recognition, some psychophysics, some computer vision. But you will see at the end the motivation was really to be able to, be able to recognize and understand really complicated things that are happening in natural images.

Now, when we look at objects in the world, people have worked a lot in object recognition, and we can recognize well complete objects, but we can also recognize a very limited configuration of objects. So we are very good at using limited information if this is what's available in order to recognize what's in there.

And this is some arbitrary collection. I guess you can recognize all of them. Some of them, if you think about a person or even a face-- this is a very small part of a face-- everybody I guess knows what it is, right? It's not even a recognizable, well-delineated part like an eye. You see a part of a person here. We know what it is, right? I mean, everybody recognizes this, and so on.

Now, I think that the ability to be able to get all the information out even from a limited region plays an important role in understanding images. And let me motivate it by one example. I'll go back to it at the end.

When we look at these images, we know what's happening. We know what the action here is. All the people are performing the same action. They are all drinking, even drinking from a bottle, right?

But the images as images are very different. If you look at each image, if you stored one image and you try to recognize another image based on the first one, it will be difficult. The variability here can be huge.

But if you focus on where the action really takes place, the most informative part where most of the answer is already given to you of what's happening here, which is where the bottle is

docked into the mouth, you can see that now these diverse images becomes virtually almost a copy of one another, almost the same. So if you manage to understand this and extract the most informative part, although it's limited and so on, the variability will be much, much reduced.

The variability here is much, much reduced compared to the variability that you have in the entire image. So most of the other stuff is much less relevant, but this is where the information is concentrated. And in the limited, restricted configuration, recognition will be much easier, and will generalize much better from one situation to another because of this principle of highly-reduced variability in the delimited image.

So we became interested. As you see, it's useful. But it's also-- to deal with small images and still recognize the limited images, you'll see it's much more-- there are some very challenging issues. And I want to discuss it a little bit, and then also discuss what it's good for a little bit more.

I will show you some human studies. What we wanted to see is what are the minimal images that people can still recognize. We examined some computational models, and I will give you-- will not keep the secret. It turns out that well-performing current schemes, including deep networks, cannot deal well with such minimal images. And, from this, I want to discuss some implications in terms of representations in our system, brain processing, and things like this.

And quite a number of people have been involved in this. Here are the names. Some of them are in the Weizmann Institute in Israel, and a few that-- Leyla is here. Leyla Isik, she is here in the summer school, and a student, Yena Han, at MIT doing some brain imaging on this, which I will mention very briefly.

So I'll start with the human study. We are looking for minimal atomic things in recognition, and the experiment goes like this. You show a subject an image and ask them to recognize it, just produce a label. So this is a dog. If they say a dog, they recognize it.

And if they recognize it correctly, we generate five descendants from this initial image. If this image was, say, 50 by 50 pixels-- and I'll tell you about pixels in a minute. But say it's 50 by 50 pixels. We make it somewhat smaller. We reduce use it, because it's still not minimal. And we reduce it in five ways.

We either copy it at one of the four corners to create, say, a 48 by 48 image by taking two

pixels from this corner here, or this corner here, and so on and so forth descendants. And we generate-- we also take the full image. Keep it as is. We do not crop it. We just reduce the resolution. So we resample it so some details start to become lost. Instead of 50 by 50 pixels, it's also a full image but 48 by 48.

And then we give each one of these images-- now we have five. We give each one-- this is beginning to expand as a tree. Each one of the five is given again to a subject. If they recognize it, again five descendants are being generated, and we explore the entire tree until we find all the sub-images which are minimal and can be recognizable in this original configuration.

Now, this is challenging psychophysically in terms of number of subjects, because we use a subject only once. Because if you show a subject this and he recognizes it, if you show him the same subject, a reduced image, he will recognize the image based on his previous exposure. So you don't you do not want to use him again.

So you don't use him again, and you show the other images to a new subject. And this requires a large number of subjects. So 15,000 subjects participated in this experiment online by Mechanical Turk, together with some laboratory controls to see that they are doing the right thing, and how it compares with the same experiment done under laboratory conditions, and so on.

So the way we define the minimal image for recognition is in this tree. Here is an image. This image is recognizable. And then we create the five descendants, and none of the descendants is recognizable. So this is recognizable. Nothing here is recognizable. So it's minimal because you can no longer reduce it either by resolution going here or by reducing the size. Any manipulation like this will make it unrecognizable.

Technically, when I'm-- in my measuring, if I'm using numbers that the image is 50 pixels or 35 pixels, and so, it's actually well-defined. I mean not the pixels on the screen. You can take the image and make it bigger or smaller on the screen. But the number of sampling points in the image is well-defined.

When you give me an image, a particular image, I can tell you how many sample points you need in order to capture this image. Technically, for those of you who know it, it's twice the-- if you do a Fourier transform and take twice the cutoff frequency, the highest frequency in the Fourier spectrum, this is, by the sampling theorem of Shannon, this is the number of points

you need in order to.

So when I said that the image was 35 pixels, I don't really care. You can make it somewhat smaller or larger on the screen by interpolation. It doesn't change the information content. It's well-defined notion mathematically how many points, discrete points or sampling points in these images.

So a very interesting thing that we found when we found these minimal images is that there is a sharp transition when you get to the level of the minimal images. So you go down and you recognize it. And then there is a sharp transition that it suddenly becomes unrecognizable, basically, to the large majority of people.

So it can change a little bit, and I'll show you some examples for you to try to see how these minimal images look like at the recognizable level, at the unrecognizable. This is the recognizable level. This is the unrecognizable level.

So to show it to you as examples here, I will show you first the unrecognizable one, the one which people find, on average, more difficult to recognize. And if you recognize it, raise your hand. Don't say what you see, because this will influence other people. Just raise your hand if you recognize the image.

And then I'll show you the more recognizable one, and let's see if more hands show up, if the distinction between the recognizable and unrecognizable holds here. I'll just show you a couple of examples from the-- OK, so I'll.

OK, so this is the one which is supposed to be difficult to recognize. If you see what it is, if you know what's the object, raise your hand. OK, good. OK, don't say what it is. We have two.

Let's see here. OK, certainly more hands. What do you see? What do you think?

AUDIENCE: Should I say it?

SHIMON ULLMAN: OK. Now you can say it because--

AUDIENCE: A horse.

SHIMON ULLMAN: A horse. Right. So let me show them side by side.

So you see that it's very difficult to recognize what's not recog-- this is, you can see the

statistic. This is recognized by 93% of the subjects. 30 subjects saw each one of these images. 93% recognized this. 3% recognized this.

And you look at the image and you see that they are very similar images, and it drops from 90% to 3%. So you can see the two images, and you can see the similarity, and you can see the large drop.

This is part of the entire tree which is being explored. This is the farther. This is the recognized one. The minimal image. And you can see that even reduce the resolution, which is really not a big manipulation, but this is a drop in performance.

And you can see all the-- so we used 50% as our criterion. So the parents should be recognized at higher. This should be recognized at lower. But typically the jump is very sharp.

Let's try two more or something just for fun. If you can recognize it, raise your hand. OK. Nobody, just for the record. OK. Look around. You can see many. What do you see?

AUDIENCE: A boat.

SHIMON ULLMAN: A boat. Right. So you can see the two images. So 80% on this. 0% here.

And you can see that what's really missing here is the tip here. And, clearly, this tip is-- there are many contours in this image, but this particular corner sharp makes an enormous difference, and it goes from 80% to 0%.

OK, let me skip. Just one more. OK, let me skip this. This is somewhat easier. OK. This is some-- OK, at least one, two, and three. OK.

How about this one? Everybody, I think. Or maybe we are missing one. So, again, you can see that the difference-- if you look at the two, there is a difference, and it's this thing here. But it's not a very big part of the image.

It's crucial, you know. You have to be trained on this. It's part of your representation. It's important. You go from almost 90% to 15%, roughly. So it's important.

So you can see that the drop is typically very, very sharp. And it's also-- the sharp transition is also interesting, in the sense that if it drops from, like the horse, from 90% to 3%, or even here, it also says that we all carry around in our head a very similar presentation.

Because if each one of us, based on the history and visual experience, would be less or more sensitive to various features, then we will not find this sharp transition. Different people will lose it at different points in the manipulation.

But at 90%, 90% of the people, roughly everybody recognizes it. You remove a feature and it goes to 3%. So everybody is using the same, or very similar, representation, which I find somewhat surprising, at least for some of these images. We don't all have the same kind of experience with horses, or with battleships, or things like that, and still the representation is very strikingly similar across individuals.

The experiment was done on 10 different objects. These are the initial objects. I showed you the object at the beginning of the hierarchy, and then you start the manipulation to discover all the minimal images inside them.

And here, so we ended up with a very nice catalog. We have a database of all the minimal images in all of these 10 images in all of the children, the unrecognizable ones. So, in terms of modeling and in terms of exploring visual features and what is necessary in order to recognize, and so on, there is a very rich data set here of all the minimal images in all of these 10 images.

Here are some more pairs of recognizable and unrecognizable. We already saw this in principle, but just to show some-- in some cases, it's pretty clear what may be going on. For example, this is horse legs, the front legs of the horse. This seems to be important.

You can see that very often it's a tremendously small-- in this fly image, very small differences, very hard to pinpoint. And it's glass that you've got in the eyeglasses. Something here is missing a little bit. But very small things in a very reliable way cause this dramatic change.

As was mentioned here, somebody mentioned, said the inflection and point, you can manipulate psychophysically a bit more. For example, here, this was another version of a minimal image. It was cropped at two locations. You can crop only the left side, or you can crop only the bottom side, and you can try to see what makes a difference. So you can really zoom in on the critical features.

In terms of number of pixels, the impression is that it's surprisingly small. So I guess you can recognize that this is an eagle. This is an airplane. And the number of pixels-- those of you who know vision, your retina has 120 million pixels. The fovea, which is the area of very high

acuity, is 2 degrees. It's about 250 by 250, 250 by 200 pixel. This is the area at the center, an area of high acuity.

But you can recognize things with, I don't know, 15, 20 pixels. It's 1/10 of your fovea. It's tiny, tiny. You can make it larger, but in terms of how much visual information, I find it surprising that you need very, very little.

It's also interesting that it's very useful, in the sense that it's very redundant. If you have the capacity, if you have a visual system that can recognize individually each one of these minimal images, and in fact they can be recognized on their own, then a full image like this contains a high number of partially overlapping minimal images. Some of them are large. You can see each one of these frame, colored frame, is a minimal image, shown not necessarily at the right resolution. You can reduce the resolution of things.

But you can see that some images are essentially low-resolution representations of the entire object, like almost the entire eagle. But some of them just contain something relatively small around the head and the eye. For the eye region, you can see that you can get a low-resolution, again, thing of almost everything. But just the corner of the eye and things like that are enough. We find, in general, it seems that things that are related to humans, you have a large number of these minimal images.

So they provide a sensitive tool to compare representations to see what's missing in the sub-image which made the image become unrecognizable. So we call them sometimes, these are called minimal recognizable configuration. We call them configuration but not images. Not parts. Not objects because they are not objects. And not parts because, as we saw in the examples, they do not have to be well-delineated parts. They are more like local configuration. But, anyway, minimal images.

The next thing that we did is, we were wondering if this kind of behavior, the ability to recognize these images from such minimal information requires-- it places an interesting challenge, or an interesting test of a recognition system, because you really have to extract and use all the available information. By definition, this is minimal. If you do not use all the information that's in this minimal image, then you don't have the minimal information. You have less than that and you will fail.

So a system that is not good enough will fail on these minimal images, or the ability to recognize them means that you really can suck out all the relevant information out. So we

were wondering what will happen if we show it to various computational algorithms that performed well on full images. What will happen when you challenge them with things which are, by nature, designed to be non-redundant?

So here is what I will do. It's not a computer vision school. I will not go too much through the details of the computational schemes, just to show you what was happening. And the bottom line is that they are not doing a good job.

Two things happen. First of all, when you train a computational system, you do not see the same drop that you see here, that it recognizes one and doesn't recognize the other. You don't have a drop in recognition. This sort of phase transition that characterizes the human vision system is not reproduced in any of the current recognition systems, including deep network and any other ones.

And, secondly, they are not very good at recognizing them. Regardless of the gap, that there is a sharp transition or not, they do not get good recognition results on these minimal images. They do not suck all the necessary information.

So in the full images, it's like we had an image of a side view of a plane. So we are training on airplane. You can think of a deep network. We actually tried a whole range of good classifiers.

And in all of these good classifiers-- those of you who are not in vision probably got enough at the beginning of this summer school that they have a feeling for a classifier in computer vision. It's a system, an algorithm, a system, a scheme, that you give it training images. You don't have to specify, you don't tell it what to look for. You just give it lots of images and tell them all these are of the same class.

And then it calibrates itself, and adjusts parameters, and so on. And then you give it new images, and the system is supposed to tell you if it's a new member of the same class or not.

So, in this case, we train the system, giving them full side views of an airplane. But then we gave them just the tails. Compared to random pictures taken from known images, the question is do they reliably can tell you that this is a tail of an airplane, part of the previous class? Or they would be confused and they will give even higher score to things which do not come from airplane at all?

So we started this when deep networks were still not the leaders, and we had some other

things, like DPM, and including HMAX, which is a very good model of the human visual system and performs very well. And so we included it as well, and deep network as well. This is the HMAX. This is convolutional neural networks.

You probably got the idea, it's just worth pointing, I find it interesting in the computer vision community that you have Olympic games every year. It's something which is very structured and very competitive, and very nice in this regard, that there is the Pascal challenge, and the ImageNet challenge, and it's well run.

And people who think that they have a better algorithm than others can submit an entry, can submit an algorithm. Everybody gets training images that are distributed publicly, but there are secret images used for testing. And you can train your algorithm on the available data. Everybody uses the same data.

And then you submit your algorithm, and the algorithm is run by the central committee on the test images. And the results are published, and everybody knows who's number one, who's number two. You have the gold medal and the silver medal. It's very competitive, and in some sense it's doing very good things. It's sort of driving the performance up.

It also has some negative effects, I think, on the way things are being done. One negative is it's very difficult to come up with an entirely new scheme which explores a completely new idea. Because, initially, before you fine tune it, it will not be at the level of the high-performing winners, and until it establishes itself as a winner, it will not get credit. So it sort of becomes a little bit conservative in this regard, which is the unfortunate part.

So, as I told you, and I will not go in great detail, the two basic outcomes is that the gap between the recognizable and recognizable-- these two bars are the gap for human vision. That's the whole group of horse images. The parents are highly recognizable. The children, the offsprings, are not recognizable. Very large drop. This drop is not recaptured in this model, in any of the model.

If you have a deep network, or you have one of these classifiers, what is recognized and not recognized depends on the threshold. You can decide that. It gives you a number, and it says that I have this and this confidence that this belongs to the class.

So what we did here is that we tried to match. We had a class of images, and people recognized them at 80% recognition. So we put the threshold in the artificial, the computer

vision system, at such a level that it recognized correctly 80% of the minimal images. So you match them.

And then we looked at how many of the sub-images passed the threshold. And you get-- this is for deep network-- that, instead of a gap, you actually got an anti-gap. It actually recognized a few more. But this should not confuse you. It does not mean that the deep network did better than humans. It actually did much worse than humans, although the bars here are higher.

And the reason is the following. You can always, even in a very bad classifier, you can get 80% recognition by just lowering the threshold, and then 80% of the class examples will exceed the threshold. The question is how many just garbage image, non-class images, will also pass the threshold at the same time. If you get 80% of the class but also lots and lots and lots of completely false positive, negative images, non-class images are also saying I'm an airplane, then that's bad performance.

So just these high bars do not say anything. The actual recognition levels were very low. We can see here for deep networks that this high bar is the performance on new airplanes. So for airplanes it did very well.

But the percent correct that it did on minimal images were 3%, or 4%, were very, very, very low. So it did very bad recognition on the minimal images. So recognition of minimal images does not emerge by training any of the existing models that I know in the world, including deep network models.

Now, the second test was, as was asked here, is that we did another large test. All of these things, actually, were a lot of effort and time-consuming. Because now we have this. This was in the original test, was a minimal image. I don't know if this was a minimal image.

Then we collected a range of tails of planes like this for many other airplanes. And we ran another Turk experiment, which was pretty large because we wanted to verify that each one of these patches that we added to our test and we were going to use for testing recognition, was indeed a minimal image for recognition.

So each one of these patches, and there were 60 of those, we ran psychophysically. And we saw that it's recognizable, and if you make it small, if you try to reduce it, it's unrecognizable. So each one of these is individually also a minimal image.

So here we did training and testing on-- so this is some examples of this. So here are various

images of fly, and each one of them was tested on 30 subjects on the Mechanical Turk. And the results are that, in terms of correct recognition, there is a substantial improvement from 3% to 60%.

But 60% is not very large. People recognized them-- I should say you should look at the false alarm. The number of errors, I will show you later. The number of errors that, even after training on minimal images, the performance of the deep network and all the other models on the minimal images is far worse than human recognition levels, human performance, on the same image.

So it's not just the gap is not reproduced. Even training with minimal images, the performance is not reproduced. The errors, or the accuracy, is far worse in all the models, including deep network, compared to human vision. So these systems do not do it.

It remains to be-- you can always ask, what happens if I train it with 100,000 images and I add and add more and more examples? This we couldn't-- this becomes more and larger.

But with the experiments we've done, which are quite extensive, it does not begin to approach human accuracy. Humans are much better. And I'll show you. I think it's not just a competition, who does better. I think there is something deeper there. And that's what I want to go next.

Let me skip some. These are the error comparison. And you can see, just as we saw, in a lot of different examples, 0 errors for humans, 17% error in the deep networks, and so on. So those are big differences.

OK. A related thing which, I think, gets to the heart of what's going on, that humans can do with these minimal images and model, at the moment cannot, is that we not only recognize these images and say this is a man, this is an eagle, this is a horse. Once we recognize it, although the image itself is sort of atomic, in the sense that you reduce it and recognition goes away, but once we recognize it we can recognize sort of subatomic particles. We can recognize things inside it.

So if this is a person, we ask again in the psychophysical test to tell us what you see inside the image using various methodologies, which I'll not go into. But people recognize this. This is a person in an Italian suit, for those of you who could not recognize it.

But once people recognize it, they say, this is the neck of the person. This is the tie. This is the

knot of the tie. This is part of the jacket, and so on and so forth. I mean, they recognize a whole lot of details, semantic internal details inside.

If they see this is the horse, the contrast is low, but they see the ear, and the other ear, and the eye, and the mouth. But if you reduce the image, they lose the recognition completely. Once they recognize it, they recognize a whole lot of structure inside.

And I think that the structure, by itself, is the more interesting part, because, really, we don't want to see a horse. We don't want to see a car. We want to know where the car door is, where the knob is. We want to recognize all the internal details.

But the ability to recognize all of these internal details is, automatically, it's also helping you with improving the recognition and rejecting sort of false detections. Because these are images the deep network thought that are good images of a man in a suit. But once you dive inside and you say, where exactly is the neck and where exactly is the tie, and is it the right structure that I expect? The answer is that it's not quite appropriate.

And you can use that so that this internal interpretation is, first of all, the more important goal of vision. But, in addition, once you do it, you can reject things that appeared, based on the causal structure, to be correct, and in this way you can get the correct recognition. And, for this reason, my prediction is that it will be very difficult to get it with current deep network, because what you'd need is not only to get the label out but to be able to dive down and get the correct interpretation, and inspect it.

And it has some properties. The tie, the knot in the tie is slightly wider than the part under it, and so on. So you have to check for the-- you know these things and you check for them. And if you don't do it, then the recognition will remain limited.

Now, when you look at it and you say, OK, and we try to develop an algorithm-- which we'll actually dive in and we'll do the internal interpretation, and we'll do them correctly and we'll reject false alarms, and so on-- it turns out that this is an interesting business. You have to be very accurate, and some of the properties and relations that you need to extract are very specific to certain categories and are very precise.

For example, this was selected by deep network as a very good example of a horse head. And, basically, it does have the right shape. But, for example, people reject it. We asked people who did not accept it as a horse head, and they said, for example, that these lines are

too straight. It looks like a man-made part rather than a part of a real animal. That was a repeating answer, for example.

But deviation, how straight is it and so on, this is a bit tricky. And also it didn't have quite the ear that you do expect here.

So we think that the kind of feature that you need in order to do this internal interpretation of interest depends on relatively complicated properties and relations that you don't want to spend time and effort doing in a bottom-up way all over the entire visual field. If certain two contours smoothly are in a corner, or if something is really straight, only semi-straight. I mean, to do all of these computations my hunch is, to do all of these complicated things, you need them only in a small-- you need some specific ones for some specific classes at some specific locations.

So the right way to do this kind of computation, the right architecture, seems to me a combination of bottom-up and top-down processing. And we know that, in the visual system-- this is a diagram of the visual system, which is supposed to show that we have lots of connections going up, but also a lot of connections going down.

And the suggestion that I would like to put up-- and I think it's what's happening here-- is that we have something like deep network that does an initial generic classification. It's bottom-up. It has some kind of-- was trained on many categories. It is not sensitive to all of these small and informative things that you need for internal classification.

And it proposes a lot of-- it gives you initial recognition, which is OK. It's especially OK when you have a complete object and not something challenging like a minimal image. Because you may be wrong on a couple of the minimal images, but you have 20 of them in each object. So if two are wrong, it's not too bad.

So, under many circumstances, you will be OK in terms of general recognition. But what this does is it doesn't complete the process, but it sort of triggers the application of something which is much more class-specific, that it says, oh, it looks like a horse. Let's check if it has, or let's now complete the interpretation.

It's not just a validation, but you really want to know where is the eye, where is the ear, where is the mouth, and so on. You want to know maybe if the mouth is open or closed. You want to feed the horse. You want to pet the horse.

I mean, when you interact with objects, all of these things are important. So you continue your understanding of the visual scene. But this is not this generic bottom-up recognition, but you are looking for specific structures that you learned about when you interacted with these objects before.

And then you test specific things. Where is the eye? There should be a round thing roughly here, and so on and so forth. So these are more extended routines that you're applying to the detected region, sort of directed from above, and you know what kind of feature to look for at different locations within the minimal image.

And this kind of ongoing, continuing interpretation is not just inside, internally, to what you succeeded to recognize, but sort of spread over the entire image. For example, if you look at this image, what do you see here in this image here? Anyone want to suggest what we see here?

AUDIENCE: A face, maybe.

SHIMON ULLMAN: Sorry?

AUDIENCE: A face.

AUDIENCE: A woman's face.

AUDIENCE: A woman's face.

SHIMON ULLMAN: A woman's face. What is the woman doing?

AUDIENCE: Drinking.

SHIMON ULLMAN: Drinking. Right. So it's a woman drinking, for those of you managed to recognize. This is the woman, and she's drinking from a cup.

Now, we tested it. The woman is actually a minimal image. If you remove the cup, you show this image, people recognize it at a relatively high. Nobody recognizes this is a glass when you just show the glass on its own.

We think that the actual recognition process in your head starts with recognizing what is recognizable on its own, sort of the minimal configuration which you know what it is. You don't need help. You don't need context. You don't need anything.

This is a woman. This is the mouth. And you can continue from there in the same way that you can recognize internally that this is the nose, and the nostril, and this is the upper lip and lower lip. In the same way that you can guide your interpretation process internally, you can also say that the thing which is docked at her mouth is a glass.

Some results from-- this has been implemented by Guy Ben-Yosef, who is also now a part of CBMM. And this internal interpretation begins to work interestingly well.

We started to do at MIT some MEG studies, because if this is correct, if the interpretation process and the correct recognition of minimal images and the following full interpretation process is driven by its-- requires for completion, it requires the triggering of top-down processing, that we could see it using the right kind of imaging. In this case, we started to do minimal images in MEG images.

MEG is-- I think you-- was MEG already mentioned here in any of the talks? So MEG, as you know, it doesn't have very good spatial resolution. It's not like fMRI, but it has very good temporal resolution. And what Leyla-- it was led by Leyla Isik.

And what we've done here is trying to let subjects in the MEG recognize minimal images. And we took the electrodes from the MEG and trained a decoder. The decoder is trained to say whether or not the image contains, say, an eagle in this case. And we had various images.

And the question is, we can follow the performance of the computational decoder that tries to say now the image, now the electrode, the pattern of electrodes, allow me to deduce that there is an eagle in the image. And we see that the decoder is successful, you can see here, at about 400 milliseconds. This is late for vision. The initial bottom-up initial recognition is more like 150, or something like this.

And we also get the same results when we do psychophysics, that in normal images you can recognize them at-- you can get good recognition after, say, exposure of 100 milliseconds followed by a mask to recognize correctly at the human level, to get to the human level that we get. With minimal images, you have to give enough time, which we suspect is enough time, to allow the application of the top-down interpretation within.

And if you don't give enough time, then people degenerate and become deep networks, and you get the same kind of performance, roughly. But this is all still unpublished and still running, and we need more subjects. And all of this is looking in the right direction, and looking in

providing support for top-down processing for this.

And this, by the way, it's interesting methodologically, because it's very difficult. With real images, it's so rich and you get so much information already in the way of going up and because of these redundancies, that even if you make 20% error, it doesn't really matter because you have redundancy, you have many multiple, sufficient minimal images within any object, and so on.

So it's very difficult to tease out the effect of where exactly the top-down information starts. Where do you need it? Where exactly you fail if you don't have it. So we think you need it for this internal interpretation and for the correct recognition of minimal images.

And here you can start seeing good signals in the MEG. It provides you sort of a tool that is pretty unique and allows you to do these things.

So let me add what is a very informal thing but where I think this is going. I think that when you look at difficult images, like action recognition that we discussed below, many things that we do depend not on sort of cause label of there is a person there, or there is an airplane, or there is a dog. But, really, things depend on the fine details of the internal interpretation.

And so if you can turn off what I think is the top-down part of class-specific top-down processes, I think that many of these fine distinctions that we make all the time-- and it's what vision is all about. Vision is not about giving cause categories-- will go away. And so these things will become more and more an important part of vision.

Let me look at this variability in action recognition. But let me show you some specific examples. This is something that confuses current classifiers, that in most of them it seems that the person is drinking. Because there is a person, there is a bottle, and the bottle is close to the mouth. So the person is drinking at this rough level of description.

But, obviously, here this person is drinking, this person is pouring, right? Something very-- is this person drinking at the moment? Yes or no?

AUDIENCE: No.

SHIMON ULLMAN: No. Why not? She's holding a cup, and it's not far, and maybe on the way to the mouth. We know that she's not drinking, right? But why exactly not?

And, again, this is something that is picked up as drinking by many recognition systems. But something is wrong here. All of these things, these are different objects and different actions that the people are performing. This is drinking from a straw. This is smoking. And this is brushing their teeth. But this depends on, you have to go to the right location and decide exactly what's happening there. It's the kind of thing that we do all the time.

Some more challenges. These are just sort of informal challenges to show you how we can deal with fine interpretation of details of interest in the image. What is this arrow pointing at?

AUDIENCE: Bottle

SHIMON ULLMAN: Sorry?

AUDIENCE: Bottle.

SHIMON ULLMAN: Yeah. But above the bottle, there is something else there.

AUDIENCE: Fingers.

SHIMON ULLMAN: Sorry?

AUDIENCE: Finger.

SHIMON ULLMAN: Fingers, right. Let's see. Just playing this time. What is this arrow pointing at?

AUDIENCE: Zipper.

SHIMON ULLMAN: Zipper. Let's see. Here are two challenging things. Here are two arrows. What is this one pointing at?

AUDIENCE: Cup?

AUDIENCE: Tea.

AUDIENCE: Cup.

SHIMON ULLMAN: All right. Next to the cup, right, is also-- this is really challenging. Let's see if some folks. What is this one pointing at?

AUDIENCE: A tray?

AUDIENCE: A tray.

SHIMON ULLMAN: Tray. So the tray, think about it, it's this, but you match it with this thing here in order to make sure, to know that it's a tray. It's not something that will be easily picked up.

I mean, I'm looking for difficult things which are a little bit challenging. And you say, ah, I can get it. But this level of detail, interpreting the fine details and images in a top-down fashion happens all the time.

Is this person smoking? Of course not, and we are not fooled by it, and we immediately zoom on the right things. And, really, all the information is here at the end of the-- and so on, and so on, and so forth.

I mean, we were looking at dealing visually with social interactions, understanding the social interactions between agents. And, again, it's very difficult to do correctly, and it depends on subtle things. I mean, you can get something rough OK.

For example, is this sort of an intimate hug, or this just a cordial hug of people who are not-- we know exactly what's going on, right? And it turns out that the features are not that easy to get. This was picked up incorrectly by something that we designed for people hugging. And it's not very far from people hugging, but it doesn't fool us, right? But they are not really hugging.

On social interactions, we know interactions even between non-human agents. I mean, this interaction, is this is threatening interaction or a friendly interaction? What do you think? Yeah. Correct. I think so too.

Anyway, I think that all of these things that we can do, and I think that vision is about this. It's not about looking at this room and saying that this is a computer and this is a chair. It's about understanding the situation and making fine judgments, and interacting with objects.

And, in fact, we're looking at is part of we're doing at CBMM. We are looking at the problem of asking questions about images. So we want a system that you can give it an image and a question, and then we want the system to be able to process the image in such a way that will give you a good answer to the question.

This is interesting because it means that it's not just a generic pipeline of running the image through a pipeline, sort of fixed sequence of operations. But, depending on what you're interested in, the whole visual process should be directed in a particular way to produce just

the relevant answer.

And we looked at a set of-- with students, we looked at a set of some 600 questions that we gave people on the Mechanical Turk images. And we say imagine some questions about these images. Ask some question about these images. And they came up with some images.

We looked at them, and an informal observation, initial observation, is that most of these questions that people invented to ask about images, you needed some things which depended on precise internal interpretation of the details. So it's things that come up all the time. You have to dive into the image and analyze the subtle cues that will tell you that these are not hugging, and this is not threatening, and this is not an intimate hug, and so on and so forth.

And this is what we are-- the whole story of the minimal images and the internal interpretation. The real goal eventually is to be able to identify the important visual features and structures which are important for this, and thinking about the automatic learning of how to extract the internal structure that will support the interpretation of all these interesting and meaningful aspects of images that, at the moment, we do not have.

OK, let me skip this. OK, I think I've said all of these conclusions already in the final comments, so let me stop here.