SHIMON ULLMAN: What I'm going to talk about today is very much related to this general issue of using vision, but with sort of the final goal of seeing and understanding, in a very deep and complete manner, what's going on around you. And let me start with just one image we've all-- we see images all the time, and we know that we can get a lot of meaning out of them. But just, you look at one image, and you get a lot of information out of it. You understand what happened before, that was some kind of a flood, and why these people are hanging out up there on the wires, and so on. So all of this, we get very quickly.

And what we do, usually, in computational vision in much of my own work is to try to understand computational schemes that I can take as an input, an image like this, and get all this information out. But what I want to talk about in the first part-- I'm going to break the afternoon into two different talks on two different topics, but they are all closely related to this ultimate goal which drives me, which is, as you will see-- I think it will become apparent as we go along-- this sort of complete and full understanding, and complicated concepts, and complicated notions that we can derive from an image.

But what I'm going to discuss in the first part is sort of how it all starts. And this combines vision with another topic which is very important in cognition-- part of CBMM as well. But it's not just this particular center, it's part of understanding cognition-- is infant learning and how it all starts. And certainly for vision, and for learning, this is a very interesting and fascinating problem.

Because here, you think about babies, they open up their eyes. And they see-- they don't-- they cannot understand the images that they see. You can think that they see pixels. And they watch the pixels. And the pixels are transforming. And they look at the world as things change around them and so on.

And what this short clip is trying to make explicit is, somehow, these pixels, over time, acquire meaning. And they understand the world, what they see. The input changes from changing

pixels and light intensities into other infants, and rooms, and whatever they see around them. So we would like to be able to understand this and to do something similar.

So to have a system, imagine that we have some kind of a system that starts without any specific world knowledge wired into it. It watches movies for, say, six months. And at the end of the six months, this system knows about the world, about people, about agents, about animals, about objects, about actions, social interactions between agents the way that infants do during the first year of life. So the goal isn't-- for me at least, it's not really-- the interest is not necessarily to engineer, the engineering part of it to really build such a system, but to think about and develop schemes that will be able to deal with it and do something similar.

I think it's also-- maybe I'll mention it at the end-- it's also a very interesting direction for artificial intelligence to think about not generating final complete systems, but generating baby systems, if you want, that have some interesting and useful initial capacities. But the rest is just, they watch the world, and they get intelligent as time goes by.

I'm going to talk initially about two particular things that we've been working on and we selected because we thought it particularly interesting. And one has to do with hands, and the other one has to do with gaze. And the reason that we selected them is, I'll show in a minute, visually, for computer vision, being able to deal with hands of people's in images and dealing with direction of gaze are a very difficult problem. And a lot of work has been done in computer vision dealing with issues related to hands and to gaze.

They're also very important for cognition in general. Again, I will discuss it a bit more later. But understanding hands and what they are doing, interacting with objects, manipulating objects, action recognition-- so hands are a part of understanding the full scheme, the whole domain of actions. And social interactions between agents is a part of it.

Gaze is also very important for understanding intentions of people, understanding interactions between people. So these are two basic-- very basic type of objects or concepts that you want to acquire that are very difficult. And the final surprising thing is they come very early in infant vision, one of the first things to be learned.

So you can see here, when I say hands are important, for example, for action recognition, I don't know if you can tell what this person is doing, for example. Any guess? Yeah, you say what? What is he doing?

AUDIENCE: Talking on the phone.

SHIMON ULLMAN: Talking on the phone. And we cannot really see the phone. But it's more where the hand is relative to the ear and so on. And you know, we can see the interactions between agents and so on. A lot depends on understanding the body posture, and in particular, the hands. So they are certainly very important for us.

I mentioned that they are very difficult to be able to automatically be able to extract them and localize them in the image. And there are two reasons for this. One is that hands are very flexible, much more so than most rigid objects we encounter in the world. So a hand does not have a typical appearance. It has so many different appearances that it's difficult to handle all of them.

And the other reason is that hands in images, although they are important, very often, there is very little information in the image showing the hand. Just because of resolution size, partial occlusion, we know where the hands are. But we can see very little of the hands. We have the impression here, when we look at this, we know what this person is doing, right? He's holding a camera and taking a picture.

But if you take the image-- you know, where the hand and the camera are-- you know, this is the camera, this is the hand, and so on-- it's really not much information. But we can use it very effectively, and similarly in the other images that you see here.

Children, or infants, do this ability to deal with hands-- and later on, we'll see, with gaze-- in a completely unsupervised manner. Nobody teaches them, look, this is a hand. Even it cannot be even theoretically possible, because this capacity to deal with, say, hands and gaze comes at the age of three months, way before language starts to develop. So all of this is entirely unsupervised, just watching things happening in an unstructured way and mastering these concepts.

And when you try to imitate this in computer vision systems, there are not too many computer vision system, learning system that can deal well with unsupervised data. And I can tell you without going-- I don't want to elaborate on the different schemes that exist and so on. But the kind of thing that exists cannot-- nobody can help, nobody can learn hands in an unsupervised way.

It may be interesting to know, just anecdotally, when we actually started to work with this, deep

networks were not exactly what they are today. If you go back in the literature, and we see when the term "deep networks" and the initial work on deep networks started by-- at least in the group of Geoff Hinton. Yann LeCun was doing independent things separately. But the goal was to learn everything in an unsupervised way. They were labeled as, that's the goal of the project, to be able to build a machine. And the machine, you will not need supervision. You just aim it at the world, and it will start to absorb information from the world and build a internal representation in a completely unsupervised way.

And they demonstrated it on simple examples-- for example, on MNIST on digits, that you don't tell the system that there are 10 digits and so on. You just show it data, and it builds a deep network that automatically divides the input into 10 different classes. And in interesting ways, it also divides subclasses. There is a closed 4, and an open 4, and so on with something very natural. And it was an example of dealing with multiple classes in an unsupervised manner.

But when you try to do something like hands, which we tried, I mean, there is just-- it basically failed as an unsupervised method. And the problem remained difficult. Here is a quote for Jitendra Malik, who is-- those of you who deal with computer vision would know the name. He's sort of a leading person in computer vision. "They say that dealing with body configuration in general, and hands in particular, is maybe the most difficult recognition problem in computer vision." I think that's probably going too far. But still, you can realize that people took it as a very difficult problem.

On the unsupervised way, which is still a big open problem, the biggest effort so far has been a paper-- it's already a few years ago-- a collaboration between Google and Stanford by Andrew Ng others in which they took images from 10 million YouTube movies. And they tried to learn whatever they could in an unsupervised way. And they designed a system that was designed to get information out in an unsupervised manner.

And basically, they managed, from all this information, to get out three concepts what happened in the machine that you-- it developed units that were sensitive, specifically, to three different categories. One was faces, the other one was cats. That's from their paper. It's not easy to see the cat here, but there is a cat. They found the cat. And there is a sort of a torso, upper body from-- upper body.

Three concepts-- after all of this training, three concepts sort of emerged. And in fact, only one

of them, faces, was really there, that there were units that were very sensitive to faces. For the other cases, like cats and upper bodies, it was not all that selective. And by the way, cats is not very surprising. You know why cats came out in these movies?

If you watched YouTube you would know, literally, it's also the most salient thing. After faces, it's literally the case that if you take random movies, or millions of movies from YouTube, many, many of them will contain cats. So in the database, it was the most-- after faces and bodies, it was the third most frequent category. So it wouldn't do, hands or gaze and so on. It's really picking up only things which are very, very salient and very, very frequent in the input.

And as I said, babies do it. And now people started to work to look more closely at it. One technique was to put a sort of webcam on infants. This is not an infant. This is a slightly older person, a toddler. But they do it with infants, and they look at what the babies are looking.

And what babies are looking at the very initial stages are faces and hands. They really like faces. And they really like hands. And they recognize hands. And they know sort of what-- they have, already, information and expectation about hands in a very early age. So it's not just even the visual recognition that they group together images of hands, but they know that hands, for example, are the causal agents that are moving objects in the world.

And this is for an experiment by Rebecca Saxe. There is a person here working with Rebecca Saxe, right? Did she talk already in the--

AUDIENCE: She will.

SHIMON ULLMAN: She will. And she's worth listening to. So this is from one of her studies in which they showed infant-- these are slightly older infants. But they showed infants, on the computer screen, a hand moving an object. This is not taken from the paper directly. This, I just drew-- but a hand moving an object-- in this case, a cup or a glass. And the infant watches it for a while and sort of gets bored.

And after they do it, they show the infant either the hand moving alone on the screen or the glass moving alone on the screen. And the hand moving alone, on itself, on the screen, they are not-- still bored. They don't look at it much. When the cup is moving on the screen, they are very surprised and interested. So they know it's the hand moving the cup. It's not the cup moving the hand or they have equal status, it's the motion of this configuration. The originator, or the actor, or the mover is the hand.

So this is at seven months. But it has been known that this kind of motion that this one-- that an object can cause another object to move, this is something that babies are sensitive to, not only at the age of seven months, but it's something that appears in infants as early as you can imagine. And you can test. And this, for us-- I'll tell you what they are sensitive to. And for us, this was sort of the guideline, or sort of the door, the open door, how to-- what may be going on that lets the infant speak up specifically on hands and quickly develop a well-developed face detect-- hand detector.

So infants are known to-- have been known to be sensitive-- they are sensitive to motion in general. They are really following moving objects. And they use motion a lot. But motion by itself is not very helpful if you want to recognize hands. It's true that hands are moving. But many other things are moving as well.

So if you take random videos from children's perspective, they see doors opening and closing. They see people moving back and forth, coming by and disappearing. Hands are just one small category of moving things. But they're also sensitive, as I said, not just to motion but to this particular situation in which an object moves, comes in contact with another object, and causes it to move.

And this is not even at the level of objects. At three months, they don't even have a well-set notion of-- they just begin to organize the world into separate objects. But you can think of just cloud of pixels if you want, moving around. They come in contact with stationary pixels and causing them to move. Whenever this happens, infants pay attention to this. They look at it preferentially. And it's known that they are sensitive to this type of motion, which called a mover event.

And a mover event is this event that I just described that something is in motion, comes in contact with a stationary item, and causes it to move. So we started with developing a very simple algorithm that, all it does, it simply looks on video, on changing images, watching for, or waiting for, mover events to occur. And the way it's defined, it's very simple. In the algorithm, we divide the visual field into small regions.

And we monitor each one of these cells in the grid for the occurrence of a mover, which means that there will be some optical flow, some motion coming into the cell, and then leaving the cell, and sort of carrying with it the content of the cell. So this does require some kind of optical flow in change detection. And all of these capacities are known that they're in place

before three months of age.

And now, look at an image of a person manipulating objects when all you do is simply monitor all locations in the image for the occurrence of this kind of a mover event. What you should see, or what you should pay attention to, is the fact that motion, by itself, is not particularly-- is not doing anything. What the algorithm is doing is, whenever it detects a mover event, it draws a square around the mover event and continues to follow it to track it for a little bit. So you can see, the minute the hand is moving something, it's detected as a hand moving things on their own, are not triggering the system here. The hand is moving. So this, by itself, is not the signal.

But the interaction between the hand and an object is something that it's detected based on these very low-level cues. It doesn't know about hands. It doesn't know about objects. But you can create-- you can see, here, a false alarm, which was interesting. You can probably understand why, and so on.

So here's an example of what if you just let it-- this scheme run on a large set of hours and hours of videos. Some of these videos do not contain anything related to people. Some of them, there are people, but just going back and forth, entering the room, leaving the room, but do not manipulate objects. Nothing specific happens in these videos. The system that is looking for these kinds of mover events finds very rare occasions in which anything interesting happens.

But you can see the output that happened. This is just examples of output of just, from these many videos, the kind of images that it extracted by being tuned specifically to the occurrence of this specific event. You can see that you get a lot of hand. These hands are the continuation. The assumption here, which we, again, took and modeled after what-- when infancy, something that starts to move, they track it for about one or two seconds. So we also tracked it for about one or two seconds. And we show some images from this tracking.

These are some false alarms. There are not very many of them. But these are examples of some false alarms, that something happened in the image that caused the scheme for the specific mover event to be triggered. And it collected these images. But on the whole, as it shows here, you get very good performance. It's actually surprising that, if you look at all the ima-- all the occurrences of hands touching objects, that 90% of them in all these collection of videos were captured by the image. And the accuracy was 65%. So there were some false

alarms, but the majority were fine.

And what you will end up with is a large collection which is mostly composed of hands. So now you have, without being supplied with external supervision-- here is a hand, here is a hand, here is a hand-- suddenly, you have a collection of 10,000 images. And most of them contain a hand. And now if you look at this, and you apply completely standard algorithms for object recognition, this is great for them-- this is sufficient for them-- in order to learn a new object.

So if you do a deep network-- but you can do even simpler algorithms of various sorts-- what you will end up using this collection of images, which were identified as belonging to the same concept by all triggering the same special event of something causing something to move, you will get a hand example. So these are just-- I will not play them-- lots and lots of movies. Some of them can play with for half an hour without having a single event of this kind. Others are pretty dense with such events. And they are being detected.

And what is shown here, it's not the greatest image. But all these squares here, the yellow squares, are-- having gone through this first round of learning these events, this is the output of a detector. You take the images that were labeled in this way. You give it to a standard computer vision classifier that takes these images and finds what's common to these images. And then on completely new images-- static images now-- this marks, in the image, following this learning, where the hands will be-- the hands are.

Now, this is a very good start in terms of being able to learn hands, and not only learn hands. As I showed you before, very early on, the notion of learn is automatically a sort of-- in terms of the cognitive system, is closely associated with moving objects, causing objects to move, as we saw in the Rebecca Saxe experiment. This also happens in this system. But this continues to develop.

Because eventually we can learn hands not only in this grasping configuration. Later on we want to be able to recognize any hand in any configuration. And so the system needs to learn more. And this is accomplished by something that, again, the details, or the specific application here, is less important than the principle. And the principle is sort of two subsystems in the cognitive system training each other. And together, by each one advising the other, they reach, together, much more than what any one of the systems would be able to reach on its own.

And the two subsystems in this case, one is the ability to recognize hands by their

appearance. I can show you just an image of a hand without the rest of it, the rest of the body or the rest of the image. And you know, by the local appearance, that this is a hand. You also know, here, you cannot even see the hands of this woman. But you know where the hands are. So you can also use the surrounding context in order to get where the hands are.

So these are two different algorithms that are known in computer vision. They are also known in infants. People have demonstrated, sort of independently before we have done our work, that infants associate the body, and even the face-- when they see a hand, they immediately look up, and they look for a face. And they are surprised if there is no face there. So they know about the association between body parts surrounding the hand and the hand itself.

And you can think about it as-- we saw this image before. Here the hand itself is not very clear. But you can get to the hand by-- if you know where the face is, for example, that you can go to the shoulder, and to the upper arm, and lower arm, and end up at the hand.

So people have used this in computer vision as well, finding hands. They also use this idea of finding the hands on their own by their own appearance or using the surrounding body configuration. And the nice thing is, instead of just thinking of, here are two methods, two schemes that can both produce the same final goal, they can also, during learning, help each other and guide each other. If you think about it, the way it goes it is shown here, that, sort of, the appearance can help finding hands by the body pose. And the body pose can do the appearance.

So if you think, for example, that, initially, I learned this particular hand in this particular appearance and pose, then I learn it by appearance. I learn it by the pose. But then if, for example, I keep the same pose but change completely the appearance of the hand, I still have the pose guiding me to the right location. And then I grab a new image and say, OK, this is still a hand, but a new appearance. Now that I have the new appearance, I can move the new appearance to a new location. So now I can recognize the hand by the appearance that I already know.

But then, this is a new pose, so I say aha. So this is still a new body configuration that ends in a hand. And then I can change the appearance again. So you can see that by having enough images, I can use the appearance to learn various poses that end up in a hand using the common appearance and vice versa. If I know the pose, I can use the same pose and deduce from the different appearances of the same hand.

And this becomes-- I will not go into the algorithms, but this becomes very powerful. And just by going through this iteration in which you start from a small subset of correct identification, but then you let these two schemes guide each other-- this kind of learning that one system guides the other-- we see, here, graphs of performance. Roughly speaking, the details are not that important. This is called a precision recall graph. But even without explaining the details of recall and precision, the higher graphs mean better performance of the system.

And this is the initial performance of what you get if you just train it using hands grabbing objects. Actually, the accuracy of the system-- the system is doing a good job at recognizing hands, but only in a limited domain of hands touching objects. So other things, it does not recognize very well. So this is shown here, that it has high accuracy, but does not cover all the range of all possible hands that you can learn.

And then without doing anything else, we just continue to watch movies. But you also integrate these two systems. Each one is supplying internal supervision to the other one. Everything grows, grows, grows. And after a while, after training with several hours of video, we get up to the green curve. And the red curve is sort of the absolute maximum you can get. This is using the best classifier we can get. And everything is completely supervised.

So on every frame, in 10,000 frames or more, you tell the system where exactly the hand is. So this is what you can get with a completely supervised scheme. And this, the green, is what you can get with reasonable training-- I mean, seven hours of training. Infants get more training-- completely unsupervised. It just happens on its own.

It's interesting, also, to think about-- if you think about infants-- and actually, I wanted to-- I was planning to ask you the question here let's see. What else could help infants do-- if you can think of other tricks in which infants somehow have to pick up, it's very important for them to pick up hands. What other signals, or tricks, or guidelines could help them pick up hands?

And OK, since I sort of gave up own hands, you can think about babies sort of waving their hands. And babies do wave their hands a lot in the air. And you can think of a scheme in which the brain knows this. And you sort of wave hands. And then the images that are generated by these motive activities interpreted by the system, it already knows, grab this, and somehow, use them in order to build hand detectors.

There is evidence that this is, I think, interesting. And we know that infants are interested in

their own hands. But there are reasons to believe that this is not the case. Because for example, if you really try this, and you try to learn hands from waving your hands in this way, imitating what infants may see, a scheme that learns hands in this way is very bad at recognizing hands manipulating objects. If, after waving the hands, you test the system, and there is a hand coming in the image and touching, grabbing something, the difference in appearance in point of view between waving your own hands and watching somebody grabbing an object is so large that it does not allow to generalize well at this stage.

And we know, from testing infants, that the first thing that they recognize well is actually other hands grabbing objects. So it's much more consistent with the idea that the guiding internal signal that helps them deal, in an unsupervised way, with this difficult task is the special event of the hand as the mover of objects.

OK, I want to move from hands to-- this will be shorter, but I want to say something about gaze. Gaze is also, as I said, interesting. It starts at about three months of age. An infant has this capability. What happens at three months of age is that an infant may look at an adult-- at a caregiver, the mother, say, or look at another person-- and if the other person is looking at an object over there, then the infant will look at the other person, and then will follow the gaze, and will look at the object that the other person is looking at.

So it's, first of all, the identification of the correct direction of gaze, but also then using it in order to get joint attention. And all of these things are known to be very important in early visual development. And psychologists, child psychologists, talk a lot about this mechanism of joint attention in which the parent, or the caregiver, and the child can get joint attention. And some people, some infants, do not have this mechanism of joint attention, being able to attend to the same event or object that the other person is attending to. And this has developmental consequences.

So it's an important aspect of communication and learning. So understanding direction of gaze is very important. And here it's even perhaps more surprising and unexpected even compared to hands. Because gaze, in some sense, doesn't really exist objectively out there in the scene. It's not an object, a yellow object, that you see. It's some kind of an imaginary vector, if you want, pointing in a particular direction based on the face features.

But it's very non-explicit. And you have, somehow, to observe it, and see it, and start extracting it from what you see, and all of this in a completely unsupervised manner. So what

would it take for a system to be able to watch things, get no supervision, and after a while, extract correctly direction of gaze?

Direction of gaze is actually depend on two types of sources of information, one of the direction of the head, and the other is the direction of the eyes in the orbit. And both of them are important. And you have to master both.

There are more recent studies of this, and more accurate studies of this, but I like this reference. Because this is from a scientific paper on the relative effect of hand-- of head orientation and eyes' orientation in gaze perception. And it's a scientific paper in 1824. So this problem was studied with experiments, and the good judgment, and so on.

The point here, by the way, is that these people look as if they are looking at different directions. But in fact, the eyes here are exactly the same. It's sort of cut and paste. It's literally the same eye region, and only the head is turning. And this is enough to cause us to perceive these two people as looking in two different directions.

In terms of inference and how this learning comes about, the head comes about first. And initially, if the caregiver, as I said, is-- the head is pointing in a particular direction but the eyes are not, the infants will follow the head direction. Later on-- so this is at three months. Later on, they combine the information from the head and from the eyes.

And the eyes are really subtle cues which we use very intuitively, very naturally. But although it's-- let me hide this for a minute. This person-- it's a bad image. It's blurred, especially those who sit in the back. But this person, is he looking, basically, roughly at you, or is he looking at the objects down there? Sorry?

AUDIENCE: audience

SHIMON ULLMAN: Yeah, basically at you, right? Now, if you look at the-- so it's from the eyes. And this is the-- these are the eyes. This is all the information. This is the right pixels, the same number of pixels that are in the image, and so on. So this is the information in the eyes. It's not a lot. And we use it effectively in order to decide where the person is-- we just look at it. And it's interesting that it's so small and inconspicuous in some objective terms. But for us, we know that the person is looking, roughly, at us.

Now, we have some computer algorithms that do gaze. And gaze, again, it's not an easy problem. And people have worked quite a lot on detection, detecting direction of gaze. And all

the schemes are highly supervised. And once it's highly supervised, you can do it. So by highly supervised, I mean that you give the system many, many images in which you give the image, but you also give the learning system-- together with the input image, you supply it with the direction of gaze. So this is the image, and this is the direction the person is looking at.

And there are ways of getting this input information of the appearance of the face and the direction of gaze, to associate them, and then when you see a new face, to recover the direction of gaze. But it really depends on a large set of supervised images. So we need something to-- if you want to go along the same direction that happened before with getting hands correctly, we want-- instead of the internal supervision, we want-- some kind of a signal that somehow can tell the baby, can tell the system-- without any explicit supervision, provide some kind of an internal teaching signal that will tell it where is the direction of gaze.

It's very close to the hand and the mover using the following notion, that if I pick up an object, which was very close to the mover event, once I picked up an object, I can do whatever I want with it. I don't have to look at it. I can manipulate it and so on. But if it's placed somewhere and I want to pick it up, nobody picks up objects like this.

I mean, when you look at objects, when you pick them up, at the moment of making the contact to pick them up, you look at them. And in fact, that's a spontaneous behavior which we checked psycho-physically. You just tell people, pick objects. You don't tell them what you are trying to do. And they're invariably always-- at the instant of grabbing the object, making the contact, they look at it.

So if we have, already, a mover detector, or sort of a hand detector, or a mover detector, that hands that are touching objects and causing them to move, all you have to do-- whenever you take an event like this, it's not only useful for hands. But once a hand is touching the object, this kind of mover event, you can freeze your video, you can take the frame. And you can know with high precision that you can-- now the direction of gaze is directed toward the object.

So we asked people to manipulate objects on the table and so on. And what we did is we ran our previous mover detector. And let me skip the movie. But whenever this kind of a detection of a hand touching an object, making initial contact with an object happened, we froze the image. Unfortunately, this is not a very good slide, so it may be difficult to see. Maybe you can see here.

So we simply drew a vector pointing from the face in the direction of the detected grabbing event. And we assumed-- we don't know-- that's an implicit, internal, imaginary supervision. Nobody checked it. But we grabbed the image, and we drew the vector to the contact point.

So now you have a system-- on the one hand, we have face images at the point where we took the contacts. So here is a face. And this is a descriptor, some way of describing the appearance of the face based on local gradients. Sorry?

AUDIENCE: How do you find the face?

SHIMON ULLMAN: You assume a face. The face detector, I just left--

AUDIENCE: [INAUDIBLE]

SHIMON ULLMAN: Right. Right. Faces come even before-- I didn't talk about faces, but faces come even beforehand. As I mentioned, the first thing that infants look at is faces. And this is even before the three months that they look at hands. This is to-- the current theory is that faces, you're born with a primitive initial face template.

There are some discussions where the face template is. There is some evidence that it may not be in the cortex. It may be in the amygdala. But there is some evidence for this face.

For manipulation of these patterns that infants look at, it's a very simple template, basically the two eyes, or something round, with two dark blobs. And this makes them fixate more on faces in a very similar way to the handling the hands. You just-- if you do this, from time to time, you will end up focusing not on a face, but on some random texture that has these two blobs or something. But if you really run it, then you will get lots and lots of face images. And then you'll develop a more refined face detector.

So babies, from day one by the way, the way we think that-- the way people think it's innate is you can work-- it is done-- experiments have been done with the first day when babies were born, the day one. They keep their eyes closed most of the time. But when they are open, they fixate. You have to make big stimuli, because it's like close-up faces, because the acuity is still not fully developed.

But you can test what they're fixating on. And they fixate specifically on faces. And once there is a face, they fixate on it. And the face can move, and they will even track it. So this is day one.

So face seems to be innate in a stronger sense. In the case of the hand, for example, as I said, you cannot even imagine building a innate hand detector because of all this variability in appearance. For the face, it seems that there is an initial face detector which gets elaborated.

So we assume that there is some kind of-- in these images, we assume that when we grab an event of contact like this, the face is known, the location of the face, the location of the contact is known. And you can draw a vector from the first to the second. And this is the direction of gaze.

And when, now, you see a new image in which there is no contact, you just have the face, and you have to decide what is the direction of gaze, you look at similar faces that you have stored in your memory. And from this stored face in memory, for this, you already know from the learning phase what is associated direction of gaze. And you retrieve it.

And this is the kind of things that you do. What we see here with the yellow arrows are collected images, which, again, the direction of gaze, the supervised direction, has been collected, or was collected, automatically, by just identifying the direction to the contact point.

These are some examples. And what this shows is just doing some psychophysics and comparing what this algorithm-- which is sort of this infant-related algorithm which just has no supervision, looking images for hands touching things, collecting direction of gaze, developing a gaze detector. So the red and the green, one is the model, the other one is human judgment on a similar situation.

And you get good agreement. I mean, it's not perfect. It's not the state of the art, but it's close to state of the art. And this is just training with some videos. I mean, this certainly does at least as well as infants. And it keeps developing, getting from here to a better and better gaze detector with reduced error. Well, the error is pretty small here too. But you can improve it. That's already-- that's more of standard additional training.

But making the first jump of being able to deal with this nonexistent gaze, collecting a lot of data without any supervision, which is quite accurate, about where the gaze is and so on, this is supplied by, again, this internal teaching signal that can come instead of any external supervision and make it unnecessary. And you can do it without the outside supervision.

It also has, I think, some-- the beginning of the more cognitive correct association, like the hand is associated with moving object, direction of gaze and going to where the-- following it to

see what is the object at the other end and so on, this is-- gaze is associated with attention of people, what they are interested in at the moment. So it's not just the fact that you connect the face with the target object and so on. It's a good way of creating internal supervision. But it also starts to, I think, create the right association that hand is associated with manipulation and goals of manipulating objects. And gaze is associated with attention, and what we are paying attention to, and so on.

So you can see that you start to have-- based on this, if you have an image like this, and you can detect hands, and you know, what-- there's a scheme that does it. You know about hands. You know about direction of gaze. You know about-- I didn't talk about it, but you also follow which objects move around. And you know which objects are movable and which objects are not movable-- so a very simple scheme that follows the chains of processing that I described.

So it already starts to know-- you know, it's not quite having this full representation in itself. But it's thought quite along the way that the two agents here. The two agents are manipulating objects. And the one of the left is actually interested in the object that the other one is holding. So you have all this, the building blocks to start build-- to start having an internal description along this line following the chain of processing that I mentioned.

And by the way, this internal training that one thing can train another, if you want, simple mover can train the hand. Mover and a hand together can train a gaze detector. It turns out that gaze is important in learning language, in disambiguating nouns and verbs when you learn language, when you acquire your first language. So this is from verb learning for a particular experiment.

But let me ignore that. A simple example would be acquiring a noun that I say. I say, suddenly, oh, look at my new blicket. And people have done experiments like this. And I can say, look at my blicket, looking at an object on the right side or looking at another object on the left side, saying exactly the same expression. And people have shown that infants exposed to this kind of situation, they automatically associate the term, the noun "blicket," with the object that has been attended to. Namely, the gaze was used in order to disambiguate the reference.

So you can see a nice-- starting with very low-level internal guiding signals of, say, moving pixels that can tell you about hands and about direction of gaze-- and then direction of gaze helps you to disambiguate the reference of words-- so these kinds of trajectories of internal supervision that can help you learn to deal with the work.

This is, to me, a part of a larger project, which we called the digital baby, in which we-- it's an extension of this. We really want to understand, what are all these various innate capacities that we are born with cognitively? And we mention, here, a number of suggested ones-- the mover, how the mover can train a gaze, and the core training of two systems. And some of the things we think that are happening innately before we begin to learn. And then we would like to be able to watch lots and lots of sensory input, which could be visual. It can be, in general, non-visual.

And from this, to generate, what will happen is the automatic generation and lots of understanding of the world, concepts like hands and intention, direction of looking, and eventually, nouns, and verbs, and so on-- so how we'll be able to do it. Know that it's very different from the less structured direction of deep networks, which are interesting and are doing wonderful things. I think that they're a very useful tool. But I think that they are not the answer to the digital baby.

They do not have the capacity to learn interesting concepts in an unsupervised way. They do not distinguish. They go, as I showed you at the very beginning, with the cats, and the upper body, and so on. They go only for the salient things. Gaze is not a salient thing.

I mean, we have internal signals that allow us to zoom in on meaningful things. Even if they are not very salient objectively in the statistical sense, there is something inside us that is tuned to it. We are born with it. And it guides us towards extracting this meaningful information, even if it's not all that salient. So all of these things are missing from the unstructured net-- or the networks which do not have all of this pre-concept and internal guidance. And I don't think that they could provide a good model for cognitive learning in this sense of the digital baby.

Although I can see a very useful role for them, for example, as just-- in answer to Doreen's question, that if you want to then get-- from all the data and the internal supervision that you provided, you want to get an accurate gaze detector, then training, using supervision training in appropriate deep networks can be a very good way to go.

I wanted to also show you, this is not directly related, but it's something impressive about the use of hands in order to understand the world, just to show you how smart infants are. I talked more about detecting the hands. It was more the visual aspect of, here is an image, show me the hand. But how they use it-- and this is at the age of about one year-- maybe 13 months, but one year of age.

Here's the experiment. I think it's a really nice experiment. This experiment was with a experimenter. This is the experimenter, one image. What happened in the experiment is that there was a sort of a lamp that you can turn the lamp on by pressing it from above. It sort of has this dome shape. That's the whites in here. You press it down, and it turns on. It shines blue light, and it's very nice. And babies like it. And they smile at it, and they jiggle, and they like this turning of the bright light.

And during the experiment, what happens is that the infant is sitting on its parent's lap. And the experimenter is on the other side, that experimenter. And she turns on the light. But she turns on the light-- instead of pressing it, as you'd expect, with her hand, she's pressing on it with her forehead. She leans forward, and she presses the lamp, this dome, and the light comes on.

And then, these are babies that can already manipulate objects. So after they see it three or four times, and they are happy seeing the light coming on, they are handed the lamp and asked to turn it on on their own. And here is the clever manipulation. For half the babies, the experimenter had her hands concealed. She didn't have her hands here, you see? No hands are under this poncho. Here it's the same, very similar thing, but the hands are visible.

Now, it turns out that the babies-- or the in-- these are not babies anymore. These are young infants-- some of them, when they were handed the lamp, they did exactly what the experimenter did. They bent over, and pressed the lamp with their forehead, and turned it on. And other children, instead of-- although that's what they saw, when they got the lamp over to their side, they turned it on by pressing it with their hand, unlike what they saw the experimenter do. Any prediction on your side what happened-- you see these two situations-- when which babies-- I mean, in this case or in this case, in which case do you think they actually did it with their hands rather than using their forehead? Any guess? Yeah?

AUDIENCE: hands in A and no hands in B

SHIMON ULLMAN: That's right. And what's your reasoning?

AUDIENCE: [INAUDIBLE].

SHIMON ULLMAN: But you think about, you know, baby, if you saw a baby, infant, young one-year-old just moving seemingly quasi-randomly and so on, something like that went on in their head, that here, she

did it with her forehead. She would have used her hands, but she couldn't, because they were concealed, and she couldn't use them. So she used her forehead, but that's not the right way to do it. I can do it differently and so on.

That's sort of-- they don't say it explicitly. They don't have language. But that's the kind of reasoning that went on. And indeed, a much larger proportion-- so this is the proportion of using their hands where the green, I think, was the hand occupied. Or when the hands of the experimenter is free, you see that there is a big difference between the two groups.

So they notice the hands. They ran through some kind of inference and reasoning. What hands are useful? Why I should do? Should I do it in the same way because that's what other people are doing? Should I do it differently? So I find it impressive.

So some general comment is-- general thoughts on learning and the combination of learning and innate structures, that there is a big sort of argument in the field, has been going on since philosophers in ancient times, whether human cognition is learned. This is nativism against empiricism, where nativism proposed that things are basically-- we are born with what is needed in order to deal with the world. And empiricism in the extreme form is that we are born with a blank slate and just a big learning machine, maybe like a deep network. And we learn everything from the contingencies in the world. So this is the empiricism versus nativism.

In these examples, in an interesting way, I think, that complex concepts were neither learned on their own nor innate-- so for example, we didn't have an innate hand detector, but also, it couldn't emerge in a purely empiricist way. But we had enough structure inside that would not be the final solution, but would be the right guidance, or the right infrastructure, to make learning possible. And this is not just a very generic learner. But in this case, the learner was informed by-- you know, was looking for some mover events or things like it.

So it's not the hands, it was the movers. And this guides the system without supervision, not only making supervision unnecessary, but also focusing the learner on meaningful representations, not necessarily just things which are the first things that jump at you statistically from the visual input. So there are these kinds of learning trajectories like the mover, hand, gaze, and reference in language, of sort of natural trajectories in which one thing leads to another and help us acquire things which would be very difficult to extract otherwise.

As I mentioned at the beginning, I think that there are some interesting possibilities for AI, as I said, to build intelligent machines by not thinking about the final intelligence system, but

thinking about baby system with the right internal capacities which will make it able to then learn. So the use of learning is sort of-- we all follow it. But the point is, probably just a big learning machine is not enough. It's really the combination of, we have to understand the kind of internal structures that allow babies to efficiently extract information from the world.

If we manage to put something like this into a baby system and let it interact with the world, then we have a much higher chance of starting to develop really intelligent systems. It's interesting, by the way, that in the regional paper by Turing, when he discusses the Turing test and how to build-- can machines think, he discusses the issue about building intelligent machines somewhere in the future. And he says that his hunch is that the really good way of building, eventually, intelligent computers, intelligent machines would be to build a baby computer, a digital baby, and let it learn rather than thinking about the final one.