HOW THE BRAIN MIGHT WORK:
THE ROLE OF INFORMATION AND LEARNING IN UNDERSTANDING AND REPLICATING INTELLIGENCE

Tomaso Poggio[1]

My research group is called Center for Biological and Computational Learning, which is in between Brain Science (Neuroscience) and the Artificial Intelligence Laboratory. We are focusing on the problem of learning. The crucial idea I would like to tell you today is the key role of learning in both understanding and replicating intelligence. We have argued, now for quite a few years, that the problem of learning is really the gateway to both understanding the brain and also to make intelligent machines. In my group we are working on the problem of learning at three different levels (see Fig.1): one is the mathematics of it, which includes quite a bit of probability theory, function approximations, and a couple of other classical branches of Mathematics.

Then there is a second level where there are the applications of what comes from the theory in terms of algorithms. These are engineering applications to a variety of domains and this is actually very interesting because it gives me an opportunity to collaborate with departments across MIT from Sloan's School, which is the Business School (people working in Finance and Marketing), to the Biology Department, the Whitehead Institute, in projects involving functional genomics, bioinformatics, and to, of course, Computer Science for projects I will describe in part, involving Computer Vision, Search Engines for images and texts on the

[1] Department of Brain and Cognitive Sciences, MIT's Artificial Intelligence Laboratory, Boston, USA - *tp@ai.mit.edu*

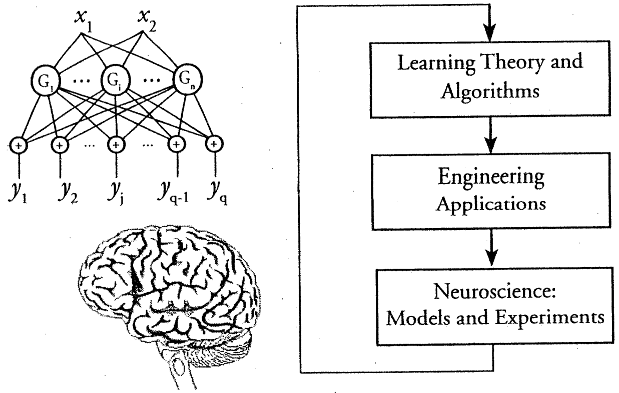


Fig.1 Three levels of the research on learning.

web and various types of data bases.

Finally, the third level at which we are working is the level of the real science in the sense of natural science. The problem we study is how does our brain learn. I will try to speak about all the three levels, even though it is not very easy to delve into the all of them.

Before I start, let me mention one explanation: why I think the problem of learning is so fundamental. If you look back to the definition of intelligence that has represented the challenge for Artificial Intelligence for the last fifty years or so, that definition was given in an implicit way by Alan Turing in the fifties under the form of the so called Turing Test. At the time that test was formulated in this way: there is a computer closed in a room, there is a person on the other side communicating with the computer through a terminal, a teletype at the time. The person can ask questions, receive answers, and this can go on for quite some time and if after a conversation of this type the person cannot decide whether he/she is speaking with a computer or with a person, then the computer would be defined as intelligent or we say that the computer has passed the Turing Test.

Already a few years ago there were some somewhat reduced versions of the Turing Test played out at the Boston Computer Museum and there were a number of programs that actually passed the test in specific domains. They were able to fool judges in believing that they were actually people. Now this did not make any news. It was not in the front page of the New York Times. This was hidden in a small paragraph in the eighth

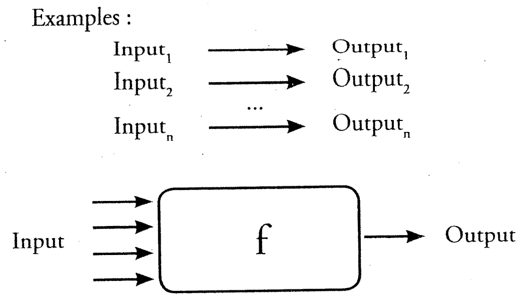
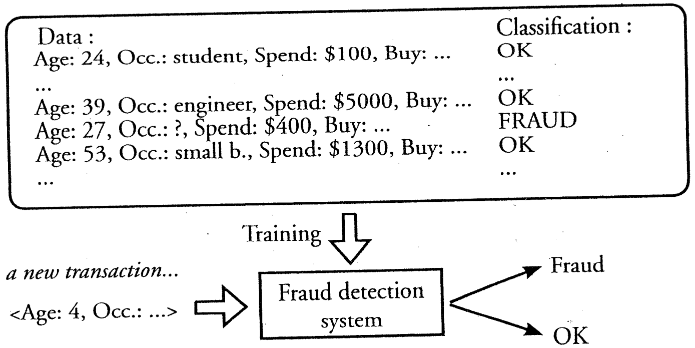

Examples :
$\text{Input}_1 \longrightarrow \text{Output}_1$
$\text{Input}_2 \longrightarrow \text{Output}_2$
...
$\text{Input}_n \longrightarrow \text{Output}_n$
Input
f
Output

Fig.2 Learning from examples.

page or so. Why is that? It is because I think the definition Turing gave in an implicit way, did not have an explicit role for learning. After all these programs were developed by experts painfully. They did not get the expertise they need by themselves, through experience, like a child does. So, I think, our present concept of intelligence requires the ability to learn from experience, and without it, I do not think we would be ready to call a machine intelligent. So this was a little story.

Let me now come to the Mathematics we have been using for formulating one form of learning, not all forms of learning, that is, supervised learning or learning from examples, which was made popular by neural networks, for instance, of which Don Norman was one of the pioneers. The idea is that you have a set of examples which are input-output pairs, the input being a vector, a set of data, for instance, various data about the weather today, say, and the output may be whether the day will be nice or rainy tomorrow. So, you have a set of data that you can collect typically from historical records, and with this set of data you train a system to try to learn this mapping from input to output in the hope that, once trained, the system will be able to make the correct predictions (see Fig.2). You can have as input one vector of data that could be interest rates in the last month of the exchange, the effect rate of euro against yen and dollar, and so on, and the desired output is whether some given market share, for example the SP500, is going up or down tomorrow.

I will not speak about this kind of application, but about another one simply because it is quite important and it will become more important very soon. This is what I see as a second wave of technologies on the world of the Internet. The first wave which is still growing and has not



Data :	Classification :
Age: 24, Occ.: student, Spend: $100, Buy: ...	OK
...	...
Age: 39, Occ.: engineer, Spend: $5000, Buy: ...	OK
Age: 27, Occ.: ?, Spend: $400, Buy: ...	FRAUD
Age: 53, Occ.: small b., Spend: $1300, Buy: ...	OK
...	...

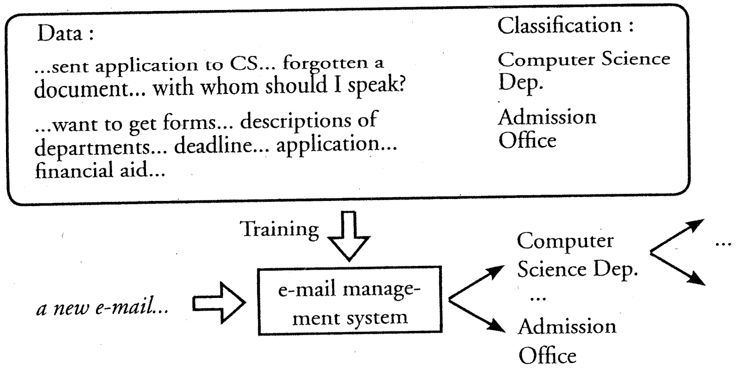
Fig.3 An automated fraud detection system.

yet reached its peek, is really about connectivity, about delivering as many bits per second as possible. This involves hardware, routers, optical fibres, companies like Cisco, Akamai. The second wave, which is just beginning, is the one of transforming all these bits into information you can use. With all this flood of bits how can you find something useful? We all know already the problems we have with the web. Search engines by all measures have become worse now than they were three years ago. There was a recent study in Nature about it. This area of technology has various names, one which I really do not like but it is commonly used, called data mining, one which is maybe better is information extraction, that is extracting information from the bits.

One technology for doing this is learning. I will tell you how.

I want first to explain the connection between this problem of extracting information and learning techniques. I will give you some examples about some projects that we have begun, then I will give a feeling of the quite deep and fascinating modern mathematics underlining what is called statistical learning. Then, I will tell you something scientific about how the brain learns. Finally, I will present you some engineering applications of learning techniques.

I will start with the topic of information extraction. This is a case which is fully used for some years by Visa and Master Card. You have a lot of characteristics of a user of a credit card, what is his age, his occupation, what are his typical spending patterns, and so on, and whether the transaction was legal or it was a fraud. You have several thousands of examples from historical records. You train a learning machine and then



Data :	Classification :
...sent application to CS... forgotten a document... with whom should I speak?	Computer Science Dep.
...want to get forms... descriptions of departments... deadline... application... financial aid...	Admission Office

Fig.4 An automated e-mail management system.

this system will tell you whether a transaction which happens right now is likely to be a fraud or not, as it is possible to see in Fig.3. There is a company in San Diego, H&C, which provides this service to Visa and Master Card. There are similar applications for telephone calls and many other things. Similar situations with customer profiling are important now for web sites like, say Amazon. You want to offer to the visitors of your site the kind of things is likely to like, possibly to buy, for instance, the books he may want to buy. Again, this is the same story. You have some data about the customer, his habits, the kinds of things he has bought and from these data, you can try to predict what the customer will do.

I will tell you about two projects we have in this area. One is a system for classifying and routing e-mail messages. We are all flooded by e-mails, companies are in an even worse situation, even the Admission Office of MIT has more than 500 e-mails per day in average and there are two people full time trying to answer these e-mails. Most of them are standard messages. They ask about admission or they have specific question for the Department of Computer Science. The idea is to train a system about e-mails and to understand the types of e-mails, see Fig.4. You can have about ten types or classes in which you can fit most of the e-mail messages. According to the category of each message you can have either an automatic prepared reply to the e-mail or route the e-mail to the appropriate person or department. We have been working on developing such a system that could classify new e-mails appropriately.

A problem that can be solved with similar techniques is the problem

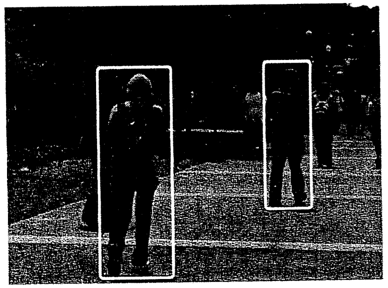


Fig.5 People detection in an image.

of finding in images people or other objects you are interested in. The idea is to train the system by showing the kind of object you are interested in. In the case represented in Fig.5, for instance, we look for people and we would like the system to find the correct objects in the given images or video sequences. The general approach in the construction of the system is to provide a training set with positive and negative examples, as shown in Fig.6. In the case of people retrieval into images or videos, the positive examples are pictures containing people, while the negative examples are images of other things, such as trees, cars, etc. After this training the system should be able to classify new images and be able to work on new data.

There are two steps in constructing such a system, see Fig.7. One is, of course, to decide how to represent the data. How do you represent an e-mail document, for instance. How do you represent an image, or a piece of text? This is the part which is like an art and very much involves prior knowledge about the domain. The second step is the one of having a learning machine which can do the desired task. This can be approached using several different techniques. It could be based on regression and the output be a real number: it may be the level of the future index tomorrow or it could be a classification, and I will speak mostly about this, in which you simply have to classify the new data as belonging to one class or another one. We are going to consider mainly binary classification, in which you are specifying just one of two classes, yes or no. For instance, whether a given portion of image is a person or not.

If we are interested in images, the first step of preprocessing the data will be the choice of some kind of representations: the values of the pixels

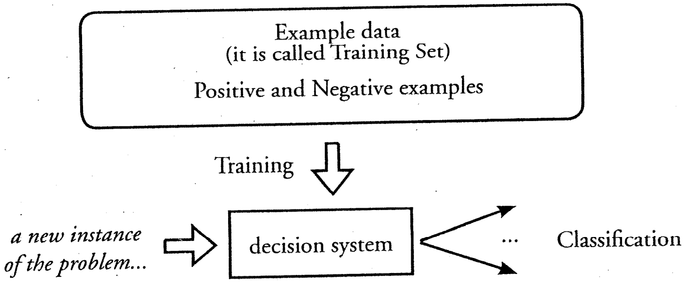

Example data
(it is called Training Set)
Positive and Negative examples
Training
a new instance of the problem...
decision system
...
Classification

Fig.6 The general schemata for training decision systems.

themselves (see Fig.8), for example. You have three colours per pixel, so an image made of ten thousand pixels will be represented with thirty thousand values. You can also have other representations such as complete or overcomplete basis of wavelets for representing the image. This is a transformation really without loss of the information contained in the image itself. The vector of data is going to be effectively what you give to the classifier. The classifier will automatically find a decision circuit, which means a surface which separates the training data into two classes: in this case people and non-people. The difficulty is that you are in many dimensions. For instance, if I use pixels as the representation method, than I am in a ten thousand dimensional space. Usually, compared to how large the space is, there are very few examples: highly dimensional spaces are very empty.

There are a number of standard approaches that have been used in the past: probabilistic methods, called Bayesian methods, neural networks, decision trees, expert systems which do not involve learning but somebody trying to extract rules from the data. We have been working on a more formal setting of the problem in which you have the set of training data: input-output data $\langle x_1, y_1\rangle$ up to $\langle x_l, y_l\rangle$ examples and you are trying to find a mapping between the input space and the output space which is also a good predictor of future values of x. I spare you the technical details, but I want to give some glimpses of what is involved. There is a theory which has been developing over the last thirty years or so, specially in the last five, which has become known as statistical learning theory and one of the many founders of it is Vladimir Vatnik a Russian mathematician who is now working at AT&T

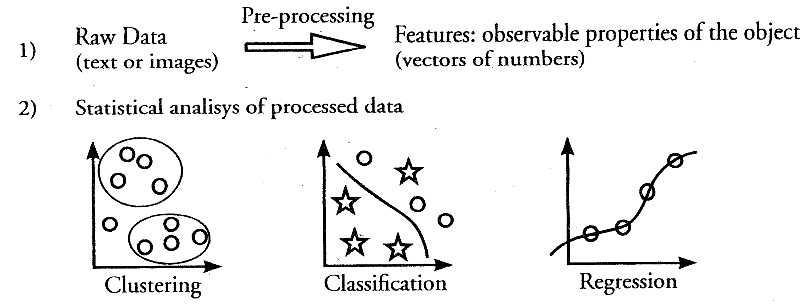


Fig.7 Features extraction phase and some approaches to decision.

Research Laboratory in New Jersey. This involves restricting the space of functions among which you are trying to find the solution. The key part of the theory is to have a space of hypotheses or mappings between input and output which has just the right complexity, not too complex, not too simple, as a function of the number of training data you have. If you have too complicated hypotheses to choose from then you are going to overfit the data: you are going to do astrology and not astronomy. If your hypotheses are too simple, you cannot capture in some cases the complexity of the mapping you are trying to learn.

Another way to look at it is presented in Fig.9 and is perhaps simpler as it makes a formal connection with a classical branch of applied Mathematics. We are given a small number of sparse data and we consider only one dimension x. The data is around the circles in the picture. For some values of x we know the output y and we would like, from these data, to be able to predict the value of y for new values of x. That is really learning. These data are the examples, and generalization simply means interpolating those data points. This is a classical problem which is solved typically using splines in one dimensions which are pieces of polynomials. If you have a linear polynomial, you have piecewise linear interpolation. The theory I am describing is really an extension of the classical approach to many dimensions. This is a framework which unifies quite a number of different approaches. It involves finding functions that are as close as possible to the data points and minimize some other properties of the function, like the smoothness of the function.

It turns out that this formulation includes a number of classical and new techniques. It is related, for example, to the theory of neural networks, in which the input data is both the input and output values, respectively x and y, of the function, and the system's output is a (guess of) the function itself, as it is represented in Fig.10. You have a number of terms and you learn the parameters related to these from the examples. Learning in this case means finding the values of these coefficients c_i in the figure. We have also constructed a software that implements this kind of learning technique.

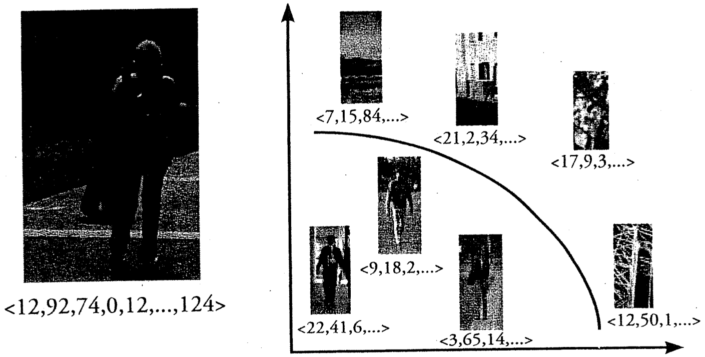


Fig.8 An example of features extraction and classification.

Let me make one more point here: the mathematics I just gave some glimpses of is quite deep. It has some even philosophical implications and this is because of theorems that establish in this theory the conditions under which something is learnable. These are really conditions about when a scientific theory is possible and when a theory can be verified. So, in a sense, learning from data is constructing a model, is being able to predict new outcomes: it is really constructing a theory. The fact that you have theorems about it is quite deep also in the sense of epistemology of science. It is another reflection of the fact that learning is a key to intelligence.

Now jumping to something which is different, and even more fundamental, let us consider how does the brain work and how does the brain learn. I will speak about one specific example of intelligence which is our ability to see and, in particular, to recognize objects, and even more in particular, to recognize a specific object, for instance, a specific face. In Fig.11 is represented the back part and the front part of the brain in different colours. Visual processes start from the back of the brain, from the area called V1, which is the primary visual cortex. There are two streams

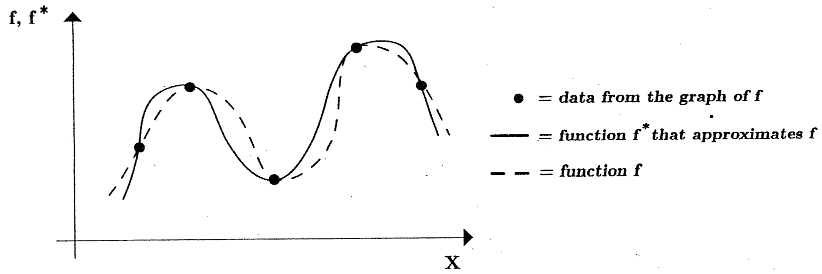


Fig.9 Multivariate function aproximation.

or flows of processes going on, one is the so called dorsal stream, which oversimplifying a lot, is involved with the question of where something is, the second one is the ventral stream, which oversimplifying a lot, is involved with which is out there, so it is involved with recognition. We have been working on this and specially in the higher visual area called infra-temporal cortex, which is known to be involved in recognition. For instance, patients who have lesions in the infra-temporal cortex have problems in recognizing objects, and recognizing faces, sometimes their own face.

We had a model quite a few years ago, which came out from the kind of mathematics I was describing, to explain how we are able to learn how to recognize a specific object from different viewpoints. The idea was quite simple. Let us consider one three dimensional object (you can think of it as a face, for example). In Fig.12 it is in fact an unfolded paper clip: there are three views of it. We have the frontal view, the side view, and one more side view. The way a learning system will learn to recognize that three dimensional object is as follows. I will describe one system in particular. You have one neuron or one processing unit which stores one of these views. Each neuron will store a different view. Then, when a new view of the object is presented to the network of neurons, after the training, the unit which stores the view which is closer to the presented view, will be most active. You can see in the figure that, by rotating the object, you may get the activity in each one of these three units going up and down depending on whether the view fits what is stored in that unit. By superimposing the output of these units, you can have an overall signal at the output of the network which is object specific, so it is high if the same object is there, but independent of the viewpoint.

$$f\,(x, y) = \sum_{i=1}^{l} c_i \mathrm{K}\,(x, y)$$

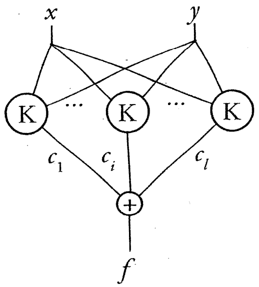


Fig.10 Function aproximation and neural netwotks.

One can test this model with computer simulation and see that it works. Additionally, there has been a collaboration with physiologists, who recorded signals from a part of infra-temporal cortex, quite from frontal quart anterior, a few years ago. The prediction of the model was that, after training the monkeys to recognize the objects like the ones shown in the picture, maybe it would be possible to recognize neurons, each one tuned to a different view of that particular object. I must say that to our surprise, that is what they have found. After the monkey has learned to recognize the given object, it was pressing the button saying yes if the learned object was presented to it, almost 100 percent correct. It turned out that each curve of the electrical activity of three individual neurons in the part of the cortex considered was having the maximum activity for one specific view of the same object that the monkey had learned. In other words, if the same happens when we learn to recognize a face, there will be neurons in our brain that will respond maximally when I see someone's face, for instance, and other ones when I see the same face from a different viewpoint. It is amazing that in our brain there are about 100 billions (10^{11}) neurons and it is likely that in these monkeys there were just a few hundreds neurons per object that the monkey had learned to recognize. Of course, there were many other neurons involved in transmitting the signal down to that part of the cortex and out of it. These few hundreds are really the ones that are storing the memory of that particular object. So the memory in this case is very localized and very specific.

We have, in the meantime, developed a more detailed model of how these neurons may work and, without giving here the details, there is a

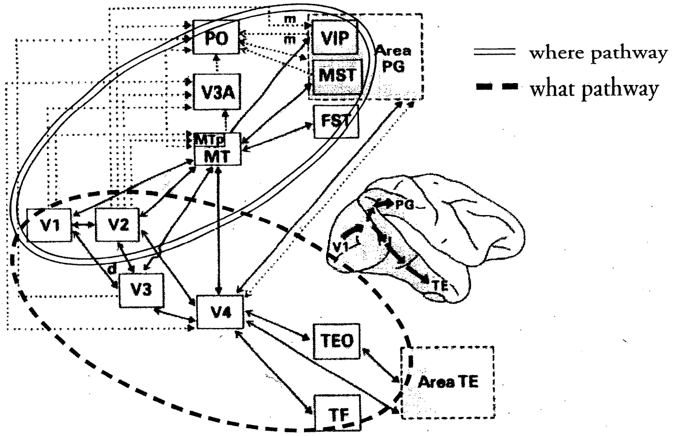


Fig.11 The "where" and "what" visual pathways in the brain.

working model that starting from the retina to the primary visual cortex gives a plausible theory of how these neurons do what they do which is really non-trivial. There are a number of experiments which seem to support the model, but it is still far from being proved. We are extending this work to deal not only with identification, as for recognizing a specific face, but with the probably more difficult problem of categorization, of categorizing a visual object as belonging to a class, for instance, to a face, or a cat, or a dog.

Again, I gave you some of the glimpses about the work that has been done on the problem of learning in my laboratory. Of course, there is much more to it. Another laboratory, next to mine, is working on the genes and the molecules that play a role in learning. They have mutant mice that cannot learn as well because one gene has been disrupted and it is affecting properties of specific neurons in a specific area of the brain, the hippo-campus. There are again various levels, from the molecular level to the circuit level, to the algorithms level at which you can study the problem of learning.

I will give you now more examples of engineering applications. First will be the e-mail system I have already mentioned to you. We have some preliminary results for that system. We used support vector machines which was one of the techniques which came out from this unified formulation I just flashed. We got pretty good results in terms of

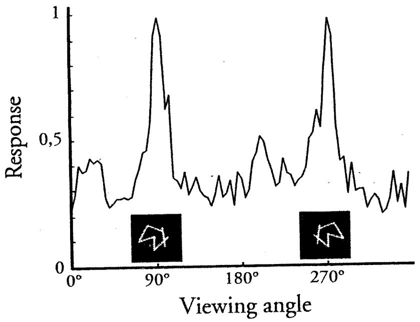


Fig.12 Activity of neurons recognizing different views of an object.

classifying data. This is classification of text which does not involve any natural language "understanding": the system is not really "understanding" the text, it is actually using statistical properties of it, like frequencies of words, frequencies of pairs of words, and some more properties of the words in the text. This is still quite far from a full "understanding" of the text. Nonetheless, with this technique it is possible to obtain classification above 80% correct in many problems similar to this.

Another engineering application of interest is the construction of systems to be trained for finding objects in images. You can think of it as a search engine working on databases of images rather than text. As I said before, the training set is made up of positive examples. It is necessary to use many hundreds of them, together with negative examples too, which are images without people. You can also train a system to find cars in images. In Fig.13 are represented some figures cars, that we have been using to train the system. When provided a new images, the system, after training, tells whether or not it contains some object we are looking for. If so, the system localizes the object: mistakes or missing detections are, of course, possible.

There is also a system which was trained to find faces (see Fig.14 for some examples). The training set was about four thousand or so positive examples and fifty thousand or so negative examples. The system was trained on real faces but it can also find a line drawing in the white board: this is a clear limitation of the system, as it demonstrates that there is no notion of context in it. As we have seen, as a consequence of the underlying model representing the dynamics of learning, the same system can be

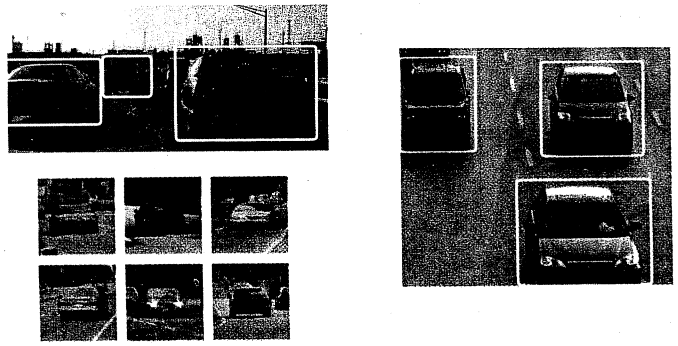


Fig.13 The car recognition system at work.

easily trained to find different objects. Therefore, despite the different training set, the system performs well.

The interest for the people recognition system was raised by the Dayton-Chrysler company, which is experimenting on a new car that, among other systems connected to a camera, has the ability to find pedestrians in order to avoid them in the downtown traffic or at least to alert the driver. Also Mercedes in Germany is doing experiments on systems working both on pictures and movies. I will spare the details of how the system works: they are not using pixels, but something called R-wavelets as a representation of the image which is then used as input to the classifier. There are some measurements about performance, detection rate, false alarms, and so on, but there is still much work to do. Finding objects in an image or a video is a very difficult problem because you have to solve it independently of context and background. For instance, in the case of people, independently on how they are dressed, different colours and contrasts and so on.

Still another application is a collaboration with a group at the Whitehead Institute which is part of MIT, this is where the largest human genoma project is going on. In this application we have used DNA chips from Afimetrics in which there are about 7100 human genes. We have data from biological tissues from patients which have cancers. We trained the system with data taken from patients of which we know which type of cancer they have. The input data are 7100 numbers which are the expression level of these genes in the biological tissues and the output is the type of cancer. Then the goal is to try to predict the type of cancer for a new patient given this signature of the patient from the DNA chip. The difficulty of the problem in this case is that there are only very few training examples. We have a high dimensional space, 7100 dimensions, we had 38 patients. However, in the problem of differentiating between two kinds of leukemia (AML and ALL), for example, we can get essentially 100 percent correct answers. In other cases, typically the ones in which we tried to predict success of a specific chemotherapy, the precision is not as high, but still quite useful for medical applications. It is quite clear that techniques of this type would probably be used in the very near future, some years from now. Again, the paradigm here is exactly the same. You have the positive examples: the data from the DNA chips for the patients you have. You train the algorithm with the data, then you make the prediction in the process. An interesting by-product in this case is that we are able to say which ones of the 7100 genes are the important ones for that particular diagnosis. We go down to about forty or fifty genes, depending on which type of diagnosis it is.

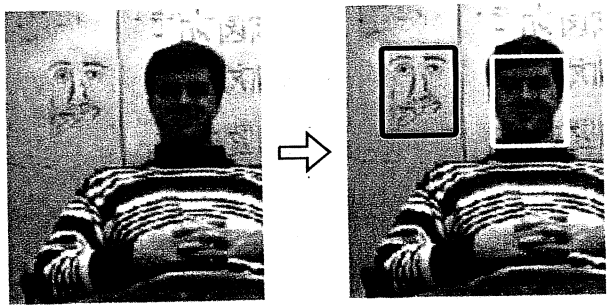


Fig.14 "Semantic" mistake of a face recognition system.

One last example of the use of learning techniques is rather unconventional. I spoke earlier about application in computer vision in which the input is an image and the output is a label or something else. You give to the learning system an image, and you what to know which object is out there and maybe also the viewpoint: this is computer vision. Suppose that you are using the same example, but now you invert the role of input and output. You train a different system in which the input is, for instance, the viewpoint and the output is the image. You may hope that, if you do things right, than the system will learn to produce images, depending on your input. What you do now is the inverse of vision, which

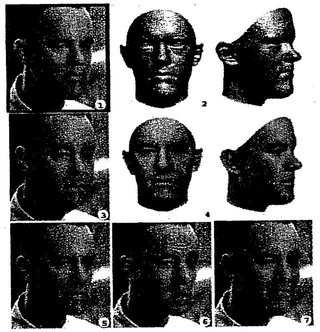


Fig.15 3D face reconstruction and synthetic images of Tom Hanks.

is graphics. In this case, you are doing graphics in a different way from the traditional one. The traditional one is really to simulate in your computer the three dimensional shapes of objects and then to simulate optics with a process called rendering, essentially simulating the paths of rays and the different object's surfaces. In the application we are talking about you are not doing this, but you are taking examples and basically interpolating among them in a multidimensional space.

These techniques has been used to construct a system called text-to-visual speech. The input is a written text and the output is synthetic speech. This is something that others have done, not us. What we have done is the visual part of it: given a few images of a person, we can make him say whatever we want. The result is completely synthetic: the person has never said those words in front of a camera. The Office of Naval Research was very interested in this work because they wanted to be able, I guess, to make movies of Saddam Hussein asking people to surrender. That is the reality.

I will show you another application of these techniques. This is work has been done at the Max Plank Institute for Biological Cybernetics in Tübingen, Germany. They have been able to take one image, the upper left of Tom Hanks in Fig.15, and from that one image, based on a system trained with about 200 faces of random people, to produce the three dimensional face of Tom Hanks. Then, when they had that, they could make various things: make him a little bit fatter or slimmer and change his expression. By using the same system trained with the same people, they can do the same thing with a now dead actress, Audrey Hepburn, and

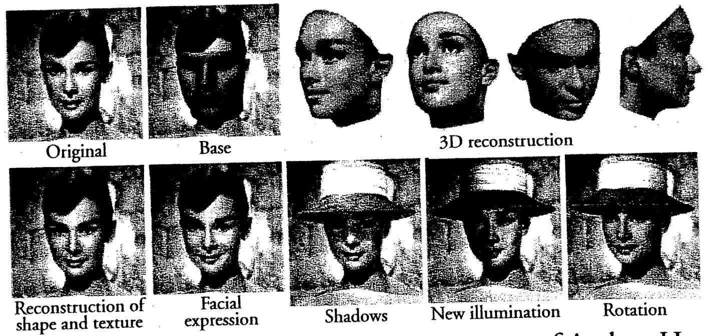


Fig.16 3D face reconstruction and synthetic images of Audrey Hepburn

take the picture on the top left of Fig.16, which is the only original one, and then obtain on the right the synthetic face produced by the system. Once they have that, they can, for instance, put an hat on her, change the direction of the illumination, change the viewpoint, and so on.

Let me conclude here. I told you about computers and brain. I told you that learning is, I believe, the key to make intelligent machines. I am also convinced that the problem of the brain and the problem of intelligence is the problem of this century and more likely of this new millennium. To be completely clear, engineering and technology are fascinating and important, also for science. However, the real science is in understanding the human brain. I think this is the greatest problem of science today, greater than the other ones which are typically mentioned, like the origin of the universe, the nature of matter, the origin of life, simply because we are using the brain to solve all problems.

One more thing to tell you: MIT has recently created a new Institute, the McGovern Institute for brain research. It is focused on studying higher brain functions. When fully staffed, the institute is going to host 300 people and 16 full-time faculty members. Their major commitment will be Neuroscience.