

# Monitor Mimics Human Facial Expressions

In the near future, advances in computer graphics and artificial intelligence may make it possible to equip your "computer colleague" with a face that would enable you to converse with it as you would with a co-worker sitting at the next desk. Key factors for the acceptance of these approaches are photo-realistic modelling and animation of "virtual" human faces. A team led by **DR. JÖRG HABER** at the **MAX PLANCK INSTITUTE FOR COMPUTER SCIENCE** in Saarbrücken, Germany, has done pioneering work in this area: the researchers developed software that is able to produce, for the most part automatically, real-time screen animations of heads that have been previously scanned. These "talking heads" demonstrate a surprisingly realistic ability to mimic human facial expressions.

## Expressions

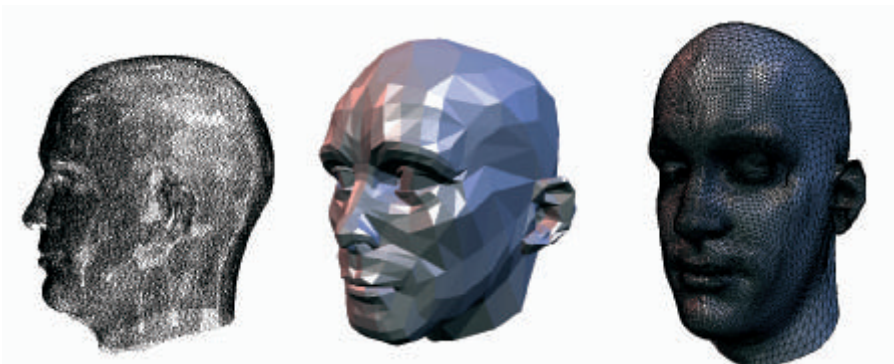


Selecting only 20 of the 200 "real" facial muscles is sufficient to give a "virtual face" the ability to imitate mouth movements required for forming sounds and making simple expressions.

**Y**es, Haber went to see *Final Fantasy* – along with his entire team. "I hope we weren't annoying the people sitting next to us with our comments. But all in all, the film was really pretty well done", says Jörg Haber, a young scientist who has been conducting research at Saarbrücken's Max Planck Institute for Computer Science for about two and a half years. But it wasn't necessarily for fun that Haber viewed this film featuring completely computer-generated actors: "Part of the reason was to see what filmmakers today are able to achieve using computer animation."

A film worth seeing, then – even though its box-office success in Germany may not have lived up to its makers' expectations. However,

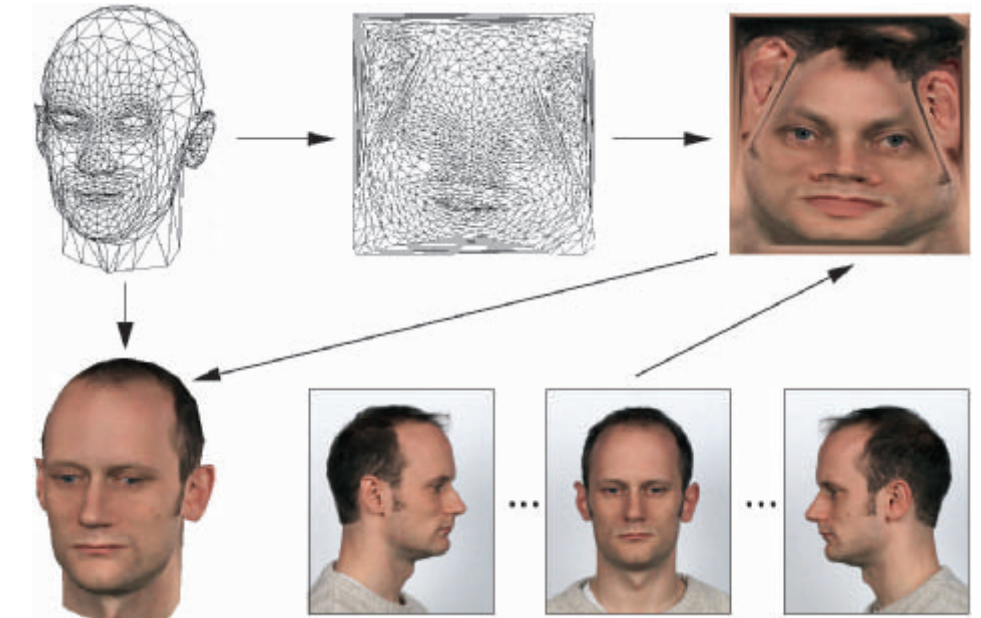
ALL ILLUSTRATIONS: MPI FOR COMPUTER SCIENCE



In order to get from scanned 3D data of a human head to a "virtual" head that can be animated, the computer adapts a generic reference head (centre) to the scanned point cloud (left). The result is an individual head model (right). This trick enables researchers to model parts of the head that the scanner may have missed.

viewers probably didn't avoid the film because of the digital characters' wooden acting skills. Indeed, the digital stars of this science fiction film seem so realistic that viewers occasionally have difficulty telling the difference between appearance and reality. But consider the tremendous amount of work required to make the film. Modelling the characters alone required the efforts of one hundred experts, about ten people per main character, toiling for roughly two years. And that was just for the preliminaries: on top of this, it took even cutting-edge graphics supercomputers up to five and a half hours to process each individual frame of the cinematic release. Haber, on the other hand, can sit back and relax: he and his team have produced software that is able to animate talking heads in real time on a typical PC – with an amazing degree of realism.

But Jörg Haber is less interested in Hollywood movies and more concerned with basic research: the question of how to make virtual heads speak, smile, and make faces with a minimum of expenditure of resources – and in such a way that it appears realistic. Haber hopes that the methods for animating virtual faces he and his team are developing in Saarbrücken will one day be useful to computer linguists and researchers working on artificial intelligence. And perhaps in the future



How to accomplish a "virtual face transplant": About three to six photos of a face – gathered from almost any angle desired – are automatically assembled by the computer and "stretched" across the previously acquired head geometry.

they will also represent "human" interfaces for Web sites or operating systems: computer programs, for example, that query data from Web users by means of a remarkably real, human counterpart that will also be able to respond to users' input with appropriate facial expressions. Haber's talking faces could thus fulfil an old dream of interface designers: the computer as a co-worker with whom one can carry on a face-to-face conversation.

### A FACE IS MORE THAN A MASK

As is apparent from the amount of effort that went into *Final Fantasy*, it is not exactly easy to imbue virtual faces with life without having them look like cold, waxen masks. Our human perception reacts in a highly sensitive manner to nuances in facial movements. Which is no wonder, considering that during the course of human evolution the ability to accurately read the expressions playing across the face of another individual

was often a matter of life and death. Things become even more complicated when one attempts to represent familiar heads, with facilities for movement and expression, in a life-like manner on a computer screen.

A face is far more than a kind of Fantomas mask stretched across a more or less rigid skull with a moveable jaw. Beneath the human complexion, when we speak, weep, or laugh, some 200 individual muscles are in constant action; some of them are simple, linear bands, while others are arranged in sheets. The orbicularis oris – the "lip muscle" responsible for forming a kissing or pouting mouth as well as the correct sounding of the "O" – is in fact a rather complex, but enormously flexible network of muscles. And all of these muscles aren't just distributed loosely under the skin; rather they affect one another during any movement.

"When you purse your lips, the muscles attached to the orbicularis oris are pulled lengthwise", Haber

explains. Even those muscle strands that are not actually actively participating in the act of speaking are deformed, and this is visible in the face: a muscle becomes thicker as it contracts, and thinner as it extends. The skin that covers the network of facial muscles is also compressed and stretched. This play of muscles and movement of skin is barely perceived at a conscious level, but has a significant influence on whether a computer-simulated facial expression will be interpreted as “real” by a human observer.

For this reason, generating photo-realistic imitations of human faces on a computer was long considered an extremely complicated undertaking – so complicated, in fact, that many animators preferred to bend the virtual faces on the screen into shape “by hand” until they arrived at a lifelike facial expression. Although there have been attempts at automatically animating human faces

since the early eighties, the successes were very limited in scope. Even many newer approaches in this research still suffer from the fact that while individual facial expressions are conveyed well, the in-between movements, such as when simulating the sound sequence “before”, appeared rather clumsy.

Jörg Haber, who had previously devoted much time to working with ray-tracing and wavelet-based image compression, found these problems intriguing. “I found it challenging to work in an area where the smallest error is immediately apparent. In addition, the simulation of faces requires – more so than in many other areas of computer science – that one brings together a whole range of knowledge from numerous different areas: geometric modelling, numerics of non-linear differential equations, generation of realistic textures, even aspects of linguistics and anatomy.” To become familiar with the expres-

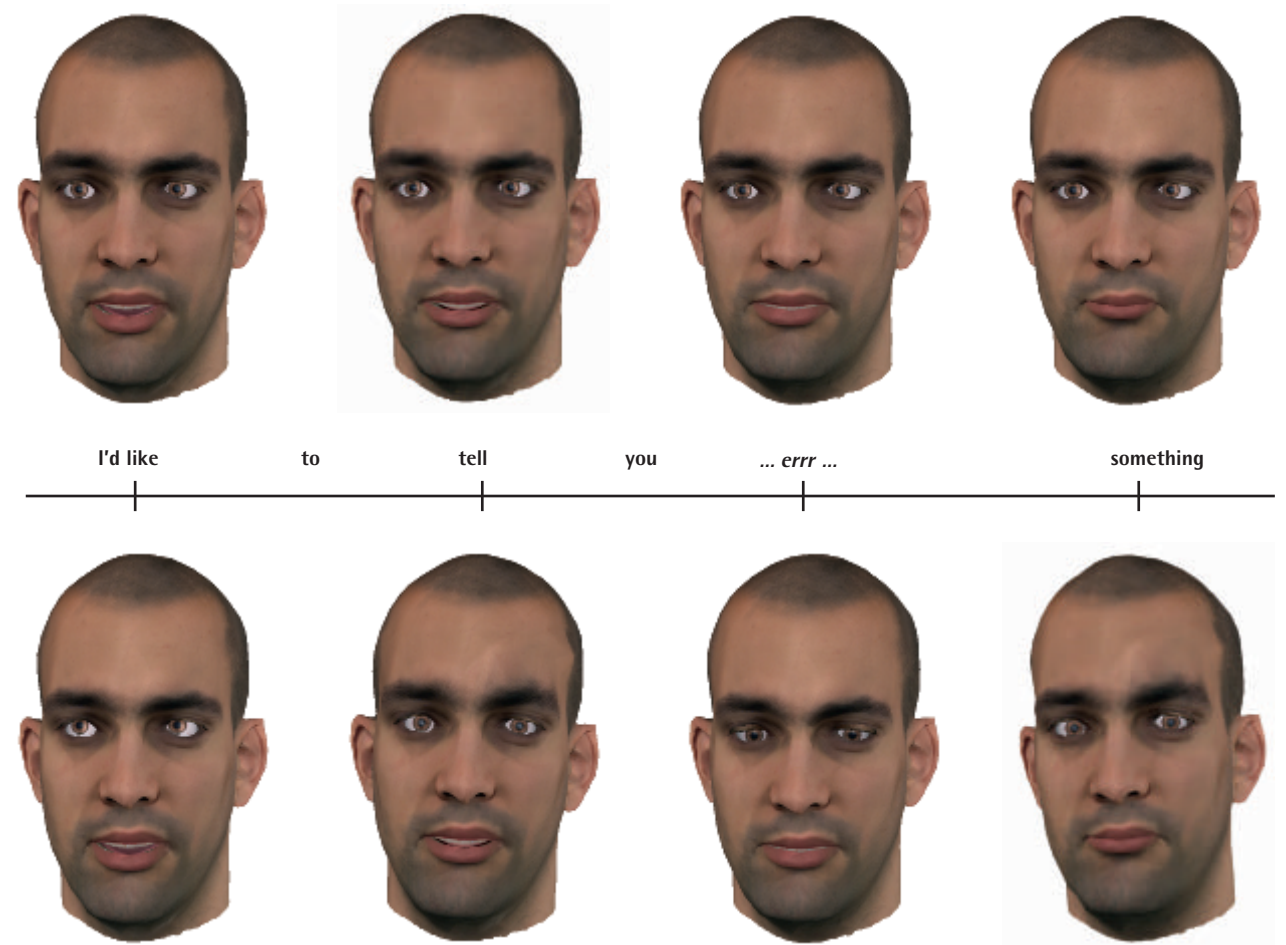
sive possibilities of human faces, Haber even consulted pertinent anatomy textbooks.

But all the work has paid off: in just under one-and-a-half years, Haber and his fellow researchers have got a handle on the task they had chosen to take on. They did this with the help of an astonishingly simple concept and real-time-capable software that can even be run on a fast personal computer. The novel idea at the heart of the Saarbrücken concept is that Jörg Haber regards faces as a wonderfully balanced aggregate composed of virtual muscles residing in a complex but manageable network of interconnected springs and mass points.

**TWENTY VIRTUAL MUSCLES FOR A VIRTUAL “O”**

These virtual faces – that in a few years may talk to you right from your screen to let you know that it’s time to defragment your hard disk once again – are at present made up of only twenty “facial muscles” that connect the elastic skin to the rigid skull according to a concept developed by Haber’s co-worker Kolja Kähler. The procedure is called geometry-based muscle modelling. The twenty muscles that were chosen can accomplish quite a lot.

“We’ve initially limited our efforts to those muscles and muscle groups that come into play in speaking and in making certain gestures, such as lifting the brow. To animate them, we’ve separated each of these into a series of ellipsoids that are deformed according to the parameter assigned to the muscle”, Haber says. The muscles in Kähler’s model become thinner when pulled, their diameter increasing when they’re contracted – with a visible result: “The chains of muscle ellipsoids are connected by virtual springs to the skin layer above and thus transfer muscle movements to the skin.”



Imitating speech-related facial expressions (below) makes a virtual counterpart more lifelike and confidence-inspiring – and even makes it easier to understand the sentences “spoken” by the computer.

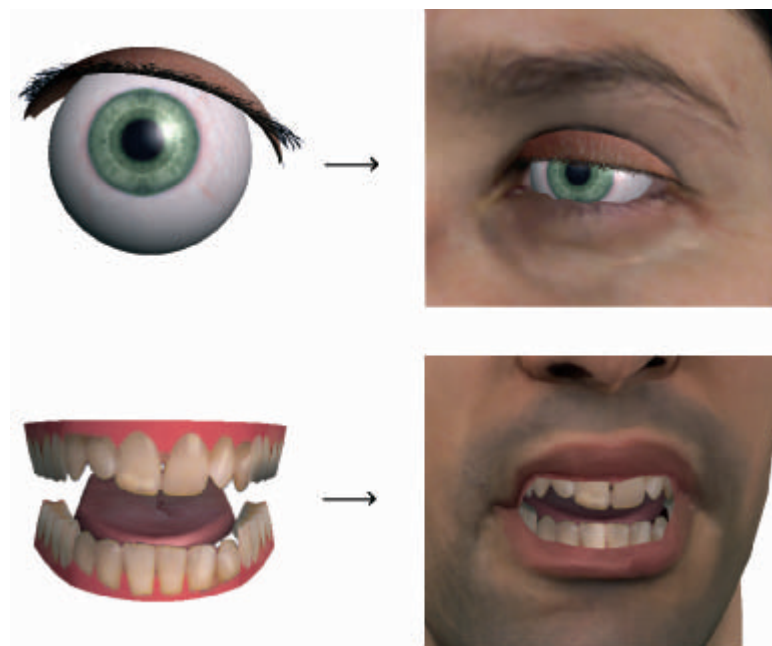
The skin itself is constructed as a dense mass spring network of several thousand links. “This enables us to simulate all the important physical parameters of skin, for instance its elasticity”, Haber says. The advantage of this approach: since the spring models that represent skin and muscles and link them together in this system are all interconnected, when one set of muscles is activated – for example when formulating an “O” – after only a few iterations of the simulation software, the entire coupled system is put into motion. Just like in real life. This results in a quite lively and astonishingly realistic facial expression. “Since we’ve limited our effort to twenty muscles, the expressiveness of these facial simulations is limited. The system is of course not yet able to express the whole range of human emotions. However, it is sufficient to communicate joy or surprise, for example.”

The Saarbrücken team has convincingly animated the extremely flexible network of lip muscles in the same way. It can pout, form an “O”, purse its lips to make a whistling sound, and it could, in principle, grip a drinking straw at any point between the upper and lower lip, if desired. And it does all this in perfect interaction with the other muscles, and of course, with the skin that covers it all. To create the movements that must be executed by the complex conglomerate of virtual muscles and skin in order to shape an “O”, for example, it was still necessary for Haber and his team to program the model “by hand”. This was complex, detailed work that the researchers accomplished with phonetics textbooks propped beside their computer monitor. “However, we on-

ly needed to do this work once, since we use the same muscle apparatus in all models, and because movements of the speech apparatus are basically the same for all people”, Haber says.

To be able to recognize a familiar person naturally requires realistically designed skin and a head shape that is as similar as possible to the original. Haber and his colleagues have developed concepts for simulating these factors as well, bundling them in a software system named “MEDUSA.” Haber was awarded the Heinz Billing Prize in November 2001 for the underlying facial modelling techniques.

What’s unique about the MEDUSA software is that, except for a few points at which it requires operator assistance, it runs completely on its own. To reproduce a “talking head,”



Correct eye colour and the right smile are important for recognizing someone you know. A single photograph of the eyes and teeth is enough to automatically generate individual models that are then fitted into the head model.



The virtual heads animated in real time in Saarbrücken are able to do more than simply form sounds: MEDUSA is also able to reproduce amazingly life-like facial expressions associated with sadness or surprise.

the program needs only a set of geometric data of the head to be modelled, along with a few photos.

First, MEDUSA takes a point cloud of a person's head generated by a 3-D-scanner, generating from it a detailed mesh of triangles. In the next step, this mesh is stretched over a virtual, idealized head model like pantyhose over a bank robber's head. In this case, however, the pantyhose does not conform to the shape of the head, but on the contrary, the computer optimises the geometry of its reference head to match as closely as possible the mesh network generated from the 3-D-data. The surface is then smoothed to the desired level of detail. Using this technique also enables the software to interpolate areas of the head that the scanner was unable to acquire, such as the back of a person's head covered by hair. MEDUSA takes special care when adapting the polygon mesh in the areas of the nose, eyes, and mouth. While it is true that the number of triangles must be minimized for faster animation, all it takes to ruin the realistic impression of the model is a single incorrect polygon.

Another advantage of the adaptable ideal head: "Since our virtual muscles are firmly affixed to this

head model, they also change position during the gradual adaptation to the geometry of the digitised person. It is thus not necessary to reposition these virtual muscles for each individual case", states Haber. Instead, the position and shape of the facial muscle fibres are automatically determined in an iterative optimization process.

**BIRTH OF A DIGITAL CLONE**

An appealing "complexion" is created by another part of the software that assembles digital photographs of a face to make a kind of rubber skin that is then pulled over the completely modelled virtual head as well; the skin is automatically and perfectly fitted. Whereas the geometric data of a head must be acquired using a 3-D-scanner, the skin can be sufficiently represented by a handful of photos shot with a digital camera from any angles desired, as long as the face is thus completely recorded. Special lighting is not required for this photo session; the posing client should simply try to maintain the same neutral expression on his face as during the acquisition of the geometric data.

To make sure that every dimple and crease ends up in the right place, however, the computer must be pro-

vided points on the head model as reference points for this digital masquerade. It takes about five minutes per photo to accomplish this manually; a virtual transplant of the entire facial skin area can be performed in 15-20 minutes. A special smoothing algorithm then compensates for any uneven colouration resulting from lighting variations in the original photographs. Finally, the person's teeth and eyes need be photographed and inserted into the image. For the montage of teeth it's enough – smile, please! – to get a shot of the front row of teeth; the remaining teeth, along with the tongue, are modelled based on default settings. After the eyes have been inserted by the computer, they sometimes need to be adjusted using a set of simple slide controls, otherwise the virtual face might appear cross-eyed.

The complicated transformation of a head into a textured polygon mesh provides a crucial advantage: It shifts the bulk of the computing time required for facial simulation so that it takes place before the actual simulation. For instance, automatic alignment of the head model to the digital "panty hose mask" takes roughly ten minutes. The resulting digital mass spring network, however, is so easily rendered on a fast dual-processor PC running a resource-efficient operating system, that it is possible to view the ani-

mated expressions of MEDUSA's virtual faces in real time, at a frame rate of about 100 frames per second. One processor interactively calculates the coupled spring network, while the other handles rendering, i.e. realistic screen output of the facial images in the correct perspective and lit in an appealing manner.

Unfortunately, for now, MEDUSA is using its extraordinary real-time facial expressions only for rather modest performances in reading pre-fabricated passages of text – in linguistics labs, for example. But who knows, perhaps MEDUSA will one day be able to read any text at all, and display the appropriate facial expression when a written passage is sad or surprising.

A first step in this direction is provided by a remarkable complement to the MEDUSA software recently introduced to the project by Haber's collaborator Irene Albrecht: a program that analyses recordings of spoken sentences with regard to their "melody" (prosody), pauses, and speech volume, then transforms the "paralinguistic information" evident in these variations into automatically generated facial expressions. Virtual heads that are made to speak in this way don't just move their lips in time with a spoken syllable, but raise their eyebrows when asking a question (which is indicated by an increase in voice frequency towards the end of a sentence), nod and wink

to emphasize something important (which can be manifested, for example, by speaking slowly and with emphasis) and even glance up at the ceiling or tilt their head when pausing and "thinking." And by making occasional, random eye movements, virtual faces also avoid a bad habit that humans find unpleasant in conversation: staring at their real counterparts all the time.

**VIRTUAL PARTNERS INSTEAD OF QUESTIONNAIRES**

The system is unable, however, to accomplish the extremely complex feat of analysing spoken sentences in real time. A fast PC still requires some six minutes to complete a paralinguistic examination of a ten-second-speech sample. On the other hand, supporting communication with the expressions of a virtual face makes it easier for humans to understand spoken language, which in turn inspires more confidence in a digital counterpart. Whether the software is at this point suited for atmospheric readings of long novels would be a matter of opinion: MEDUSA is not yet able to recognize and provide appropriate facial expressions for various shades of intonation that convey fear, joy or boredom. So, it will likely be some time before grandmothers can get a break from reading bedtime stories aloud. The effort required for thinking – the automatic "decoding" and com-

prehension of a text that is available only in printed form – is still far too complex.

We may, however, conceivably encounter MEDUSA's faces on the World Wide Web in the not-too-distant future: "In the foreseeable future, it should in principle be possible to make software that can animate a head available as an applet online, from where it can be downloaded and run on a home PC", Haber opines. Perhaps over-the-top virtual heads such as Max Headroom or Lara Croft will soon be substituted by real-looking characters appearing in our Web browsers, engaging us in personal conversation to compel us to buy detergent and other products.

And what if one day George Lucas or some other Hollywood producer calls in search of support for a new virtual flick in the vein of *Final Fantasy*? "I'd be happy to explain our techniques to him, but we're not interested in providing services for film production because to imitate actors in a way that is truly realistic, we'd need to animate not just 20 but all 200 facial muscles", Haber says. "Although it would be possible at this point, that would just be a laborious task without any scientific appeal for us. Before working on something like that, we'd rather turn to new and more interesting topics." Which leaves the rest of us to eagerly anticipate what those might be.

STEFAN ALBUS



A series of images from a sequence generated by MEDUSA in real time, in which only those muscles required for producing sound were animated. Since all people use their facial muscles in a similar way when speaking, the correct movement sequences for forming sounds needed to be worked out only once. Once these have been established, they can easily be transferred to all other scanned heads.