Computers **Make Faces**

These days, animated figures in films and in computer games are often true to life. After all, they are created with sophisticated three-dimensional models of bodies and faces. **Christian Theobalt** and his colleagues at the **Max Planck Institute for Informatics** in Saarbrücken are making it much easier for graphic artists to generate such models – enabling applications that were previously inconceivable.

TEXT **TIM SCHRÖDER**

Actor Tom Cruise is a cool guy. In Minority Report, he shines as a hard-boiled fighter against evil. He is cast as a policeman in the year 2054 – and as you would expect from science fiction, he is surrounded by a huge amount of high tech. But looking at the movie today, only 14 years after its premiere, there are parts of the vision of the future that don't look very advanced. In a scene that seemed very futuristic back then, the actor moves his hand to open screen windows on a glass wall that serves as a monitor. He stretches and shrinks the images with finger motions, then conjures them away with a brisk swipe. That looks stylish, but when he does so he is wearing a black glove with luminous dots – a data glove. And today scientists simply grin at that.

This principle, where Tom Cruise controls a computer with a data glove, has been around in movies for about 20 years, and is still being used for new productions. A person is filmed wearing a whole bodysuit with markers that a computer can use to follow – track – the position of the head, body, arms and legs. This enables the movement to be transferred into a scene in a movie of a computer world, for example, to animate a fictional figure in a human way. But this kind of tracking is complex. Actors have to struggle into gloves and a full bodysuit.

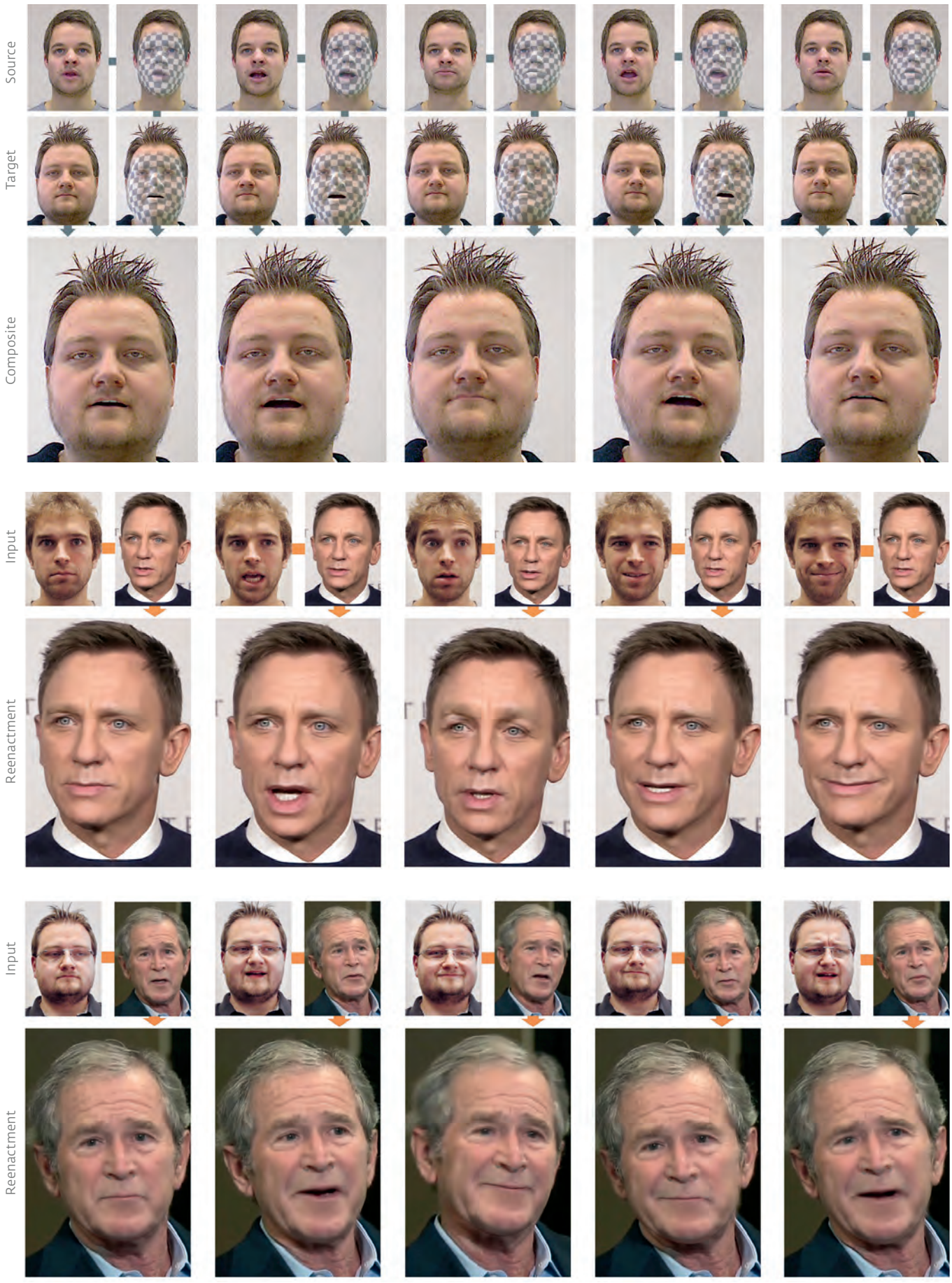
ANALYZING REALITY IN MOTION WITH THE LEAST EFFORT

Christian Theobalt and his colleagues at the Max Planck Institute for Informatics can take some of the credit for the fact that tracking with markers is gradually going out of fashion. The computer scientist leads the Graphics, Vision and Video Research Group at the Institute. He is working at the interface between computer graphics and image recognition. Theobalt's main aim is to teach computers to analyze reality in motion with the least possible effortand, above all, at high speed, and then convert this into accurately detailed three dimensional virtual models that precisely capture the shape and reflective properties of a figure, as well as the lighting of a scene.

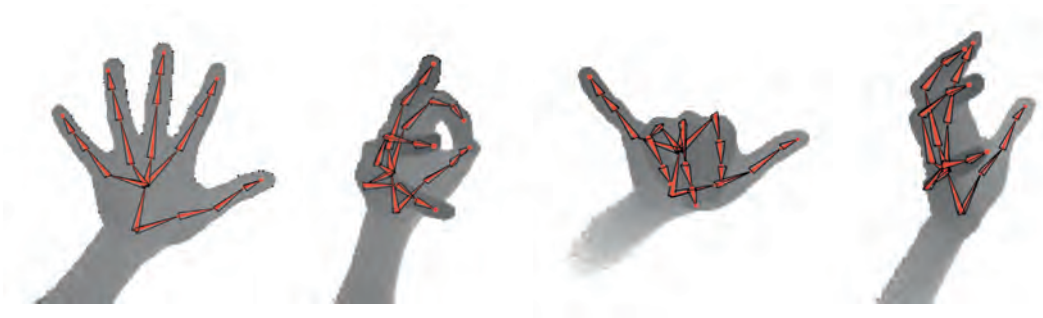
His working group has two main research focuses: the recognition of movements and the analysis of faces. It isn't only filmmakers who could use these methods, to adjust the expression of an actor or animate fantasy figures. Computer game developers could similarly provide avatars with natural movements and facial expressions. And robots could utilize the techniques of scene analysis so that they could get an exact picture of their environment. Above all, they must be able to distinguish people in the background from relevant objects in the foreground. Another major application area is augmented reality, the mixing of reality with a virtual world to allow virtual objects to be moved with gestures.

In the future, Theobalt's research should save the film and games industries alone thousands of person-years in work throughout the world, since it isn't only the tracking with markers

Right Changing faces: Computer scientists from Saarbrücken transfer the facial expression of an actor in a source video onto an actor in a target video. Gray and brown arrows indicate the workflow. In the top two rows, the morphable 3D face models of the source and the target are shown in a chessboard pattern laid over the tracked faces. In the fourth and sixth rows from the top, the original facial expressions of the source and the target can be seen.



Photos: MPI for Informatics



Franziska Müller (right page) has developed a software that efficiently tracks even subtle movements of a hand with rather low calculating effort and reconstructs the gestures in a 3D model. To make sure that the captured positions are anatomically possible, the program utilizes a model of the hand skeleton, shown in red (left page).

that makes the animation of films and video sequences an arduous task. Graphic artists convert recordings of a real person into a mathematical model, generate figures from it and transfer these into a computer game or movie sequence.

Today, a lot of this work has to be done by hand. For example, to get the artificial face of an actor that is copied into a particularly spectacular move sequence to smile convincingly or to wrinkle its forehead requires computer graphic artists to invest many hours of precision work.

As far as the first part of his research goes – motion capturing – there are still some obstacles to overcome. For example, in order to allow the scary Gollum to appear in the Hobbit and Lord of the Rings movies, the filmmakers transferred the body and the movements of an actor onto a model. They adapted the pose and in particular the face to their idea of the fantasy creature, textured the figure with Gollum skin and copied it into a cave world generated on the computer.

In order for the actor's movements to be converted into a model by the computer, the actor, and particularly the face, must be perfectly illuminated in special studios and recorded in different poses by several cameras so that concealed parts of the body are captured. And of course the recording studios must be relatively empty apart from the actor, as any objects present would confuse the computer in the analysis of the scene.

"It would be much simpler if we could just record people in the open air, in a perfectly normal environment and with varying lighting conditions," says Theobalt. "And preferably with only one camera to reduce the effort – that's precisely what we are aiming for." Until recently, that was inconceivable. Step by step, Theobalt and his ten colleagues are approaching their target.

RECOGNIZING THE MOST SUBTLE GRASPING MOVEMENTS

Doctoral student Franziska Müller and her colleague Srinath Sridhar want to help the computer recognize the movements of hands more reliably and transfer them to a three-dimensional model. That is particularly difficult, because the rapid movement of a hand and its fingers often leaves some parts out of the camera's field of view. But being able to follow the fingers exactly is important for operating devices with gestures in augmented reality. "For this, the computer must be able to interpret the hand gestures correctly," says Franziska Müller.

Müller is already much further than what Tom Cruise did in Minority Report. Cruise rearranged the images on a monitor with coarse hand movements. But Müller's computer can recognize the most subtle grasping movements. To do this, she switches on a small camera on her computer screen and measures the three-dimensional shape of an object with laser beams. On the otherwise white screen, an artificial hand appears and follows all of her finger movements. Müller presses her thumb and index finger together. She opens and closes her hand. And the artificial hand on the screen carries out every movement.

As so often, the devil is in the details: the computer must keep calculating the position of the fingers, and that in fractions of a second, or else the image on the screen would falter and jerk. And of course analyzing the gesture must take into account the parts of the hand that happen to be concealed from the camera. "That's only possible with mathematical procedures that reduce the quantity of image data and can still calculate the position correctly," explains Franziska Müller.

Concretely, it is a matter of the mathematical analysis of distance data. The small depth camera on Müller's monitor measures the run time of the light for every image point – for instance, to a fingertip or the ball of the thumb and back.

Franziska Müller switches on another program she bought from a software company that already offers a program for real-time hand measurement. The result is disappointing: When the researcher moves her hand quickly, the program can no longer follow. The model of the hand on the monitor suddenly loses fingers; for a moment, a finger appears in a wrong position. It gets really bad when one finger conceals another. The hand on the screen partially dissolves. The software she bought has problems importing the image data correctly into the model.

The reason: conventional programs can't handle the enormous computing load for calculating stable three-dimensional movements from images from one camera perspective. Müller therefore uses a different procedure. Her software arranges the measured values for individual pixels so that neighboring pixels at the same distance from the camera are represented as Gaussian

Graphic: MPI for Informatics

Photo: Oliver Dietze

clouds. In this way, the number of points can be reduced considerably. This shortens the computing time, which allows Müller's program to keep up even when she moves her hand quickly.

Müller's software compares the calculations from the distance measurements with a model skeleton already stored in the program. This gives the computer an idea which hand and finger positions are possible.

Müller also uses a computer learning procedure that estimates which part of the hand a pixel belongs to in fractions of a second on the basis of probabilities. Franziska Müller has fed the computer with training data: it has learned how a hand can look when it is rotated or moved. In addition, she has built in another kind of error estimation in her program that excludes values that make no sense in terms of hand anatomy.

"Thanks to Franziska's work, we can now measure subtle finger movements too, such as when one rubs thumb and index finger together," says Theobalt. "Conventional programs can't resolve that at all."

The hand is, of course, not everything. In many cases, the movement of a whole body must be captured. The Saabrücken-based team uses a skeleton model that was developed in Theobalt's group and gives their software a degree of anatomical knowledge.

"We are slowly departing from classical motion analysis," says Theobalt. The computer normally orients itself toward characteristic structures in a sequence of images, not only marker points, but also image regions that have a similar appearance. "We call this procedure correspondence finding. The computer tries to follow an object that slowly moves in a sequence of images." The problem: with changes in lighting, these procedures produce significantly more errors, because the corresponding image points keep changing their luminosity.

Theobalt's team has not only made the movement analysis more independent of the environment and lighting, but reduced the number of cameras necessary from over eight to three. For this purpose, Theobalt applies machine learning. In this way, the researchers can compensate for the computer's briefly losing track of concealed body parts with only a few cameras and changing lighting. They train the machine learning tool with images of different poses so that it learns to identify the body parts.

The combined approach makes the movement analysis of Theobalt's group especially efficient. This is the first procedure that can measure the movement of the complete skeleton in 3D quickly and very robustly – not just in a carefully lit studio, but outside in any kind of environment with continually changing lighting conditions.

Former doctoral students and postdocs of Theobalt's have founded the company TheCaptury, which specializes in motion analysis with the help of the skeleton model. The company offers software that analyzes the position and motion of the limbs from video recordings from one or a few cameras, even in real time. "The software is used to analyze the fast motion sequences of athletes or to investigate the body positions of people at their workplace," says Theobalt.

The challenges with facial recognition, the second focus of the working group in Saarbrücken, are altogether similar. To produce realistic-looking high-resolution 3D facial models, the face of a person must be illuminated in a defined way and recorded by several

cameras. Only then can the computer calculate the three-dimensional shape of the face and cleanly reconstruct little wrinkles as well as reflections from the skin. To transfer the face of an actor into artificial worlds, they have to record many different facial expressions: laughing, looking mean or raising their eyebrows. It takes a lot of effort to model an expression that hasn't been previously recorded onto a face in a film scene.

CONSTRUCTION OF THE MODEL IN FOUR PARALLEL STEPS

The role of facial recognition is not limited to the movie and computer game industries. Novel fatigue alarms in cars rely on the interpretation of facial features, for example. Some companies are also working on procedures for interpreting lip movements. For instance, automatic speech recognition could be significantly improved, as not only the audio channel is used, but also the lip movements in the video image.

Theobalt wants to simplify facial recognition in a similar way to motion analysis and create three-dimensional models that can produce facial expressions that were not recorded in the creation of the model. His team is working on transferring video recordings made of faces with a single camera and uncontrolled lighting into 3D facial models. Unlike conventional calculation procedures, the technology is so fast that the model can follow the expressions of a filmed person.

In order to reconstruct a moving artificial face in acceptable time from the simple video image of a single camera, Theobalt must travel a different path from previous methods. He calls it inverse rendering. The term rendering, from computer graphics, stands for the precise calculation of correctly illuminated images from a model of the scene. In inverse rendering this is turned around, and the model of lighting, reflectivity and geometry that best elucidates the appearance and the shading in the image is calculated. The facial reconstruction becomes very robust to scene changes and functions independently of whether the sun is shining or the sky in front of the window is overcast.

The trick: Instead of analyzing a face with wrinkles, shadows and reflections under studio conditions pixel by pixel, Theobalt's team divides the construction of the model into four parallel steps: first, the recognition of the shape of the face; second, the reconstruction of how this changes with different facial expressions; third, the estimation of the reflective properties of the surface of the face, known as reflectivity; and fourth, the estimation of the lighting in the room.

The challenge associated with the recognition of the shape of the face and its changes is to extract spatial information from the two-dimensional video signal from the camera – the position of the prominent nose or sunken eyes, or the shape of the mouth. "We overlay the image of the face with a 3D facial model that was developed here at the Institute a few years ago (see MaxPlanckResearch 4/2011, p. 62)," says Theobalt. "Its strength is reconstructing a 3D face from sparse image information."

Separately from the recognition of the shape, inverse rendering analyzes the information about light and shadow in a scene, or the reflectivity. "First we calculate from that the lighting conditions prevailing in the space," explains Theobalt. Subsequently, the computer can let the light in the space react with the shape of the face. It can then draw conclusions about the 3D shape, including fine details, from the shading in the face in the video. In several iterations, the computer compares the facial model it produced in fractions of a second to resemble the actual video image, and adjusts it until it agrees with the original. This happens so fast that the model can smoothly reflect even rapid changes of the features of the face.

How well inverse rendering works was recently shown by Theobalt's postdoc Michael Zollhöfer together with colleagues from the universities in Stanford and Erlangen. The researcher made a splash in the media when he succeeded in transferring the expression of one face to another in real time – "reenactment" of an expression.

Zollhöfer shows how this works. He switches on a conventional camera the size of a bar of chocolate and photographs his face. It appears on the monitor, which the computer first covers with

The method of Christian Theobalt's team also captures the motion of whole bodies in 3D models, shown here for a boxer.

Graphic: MPI for Informatics

To analyze reality in motion and to capture not only the shapes of bodies and faces, but also their reflective properties and the lighting of a scene is the mission of Michael Zollhöfer, Franziska Müller, Abhimitra Meka, Dushyant Mehta, Hyeongwoo Kim, Pablo Garrido and Christian Theobalt.

a grid. "The computer is now calculating the model of my face, which takes a few seconds," says Zollhöfer. But then it goes quickly. Like a Venetian carnival mask, the three-dimensional animated representation of Zollhöfer's face appears on a second screen. If Zollhöfer moves his mouth, the mask follows his movement.

Then he switches on a video of Arnold Schwarzenegger being interviewed on a second monitor. The software generates a model from Schwarzenegger's face in the computer. Then the sensation: When Zollhöfer opens his mouth, Arnold also opens his mouth. Zollhöfer wrinkles his nose, grins, wrinkles his forehead – and the image of Arnold obediently follows every motion. "As you can see, my facial expression is transferred in real time onto the facial model of Arnold Schwarzenegger," says Zollhöfer.

For the movie industry, this means that a naturally appearing expression can be directly transferred from one person into the moving video sequence. This is a minor revolution. Not least because one can now simply transfer any kind of facial expression into a model of a face.

Some producers have since already knocked on his door. However, he still has to disappoint them. "We still have to optimize our facial model, especially the lip movements, because people are extremely good at noticing small inaccuracies," says the researcher. If the lips don't close one hundred percent correctly with a sound, that has a very disturbing effect. "But I think we'll get there in a few years," says Theobalt. ◂

TO THE POINT

- **In order to achieve natural effects for the expressions and movements of animated figures in movies, computer games or other applications in virtual or augmented reality, graphic designers have been using three-dimensional models of faces or bodies generated with enormous effort.**
- **Christian Theobalt and his colleagues are developing methods to analyze and transfer the movements of faces and bodies to models using recordings from one or a few cameras with arbitrary lighting and in an arbitrary environment, and with relatively little computer power.**
- **The researchers can use anatomical models stored in their software, as well as methods of computer learning.**
- **Thanks to the minimum effort required to transfer movements into three-dimensional models, applications that used to be inconceivable will now become possible. For example, the researchers can transfer the expression of one person onto the face of another in real time.**

GLOSSARY

Motion analysis: Various methods are used in this technology, also known as motion capturing, to record the movements of people in three dimensions. Older procedures here depend on markers and precisely defined recording conditions.

Computer learning: Computers are trained for various tasks using many data sets. This helps them learn to recognize objects such as a table, even if they have seen only similar objects in the past, or see the object from an unfamiliar perspective.

Photo: Oliver Dietze