Photo: MPI for Informatics

Digital **Storytellers**

Movies with audio descriptions help blind people understand the storyline. Could computers take over the task of transforming moving images into natural language? **Anna Rohrbach**, a scientist at the **Max Planck Institute for Informatics** in Saarbrücken, and her husband, **Marcus Rohrbach**, who conducted research at the same Institute until recently, have made it their mission to make that possible. They aim to develop a computer that can automatically generate and read out film descriptions.

TEXT **TIM SCHRÖDER**

The Pianist," "Gandhi," "Men in Black," "X-Men" – Anna Rohrbach owns a sizeable collection of movies and blockbusters. Her office shelf is filled with around 200 DVDs, neatly sorted in rows. While most people collect DVDs with the intent of spending a cozy movie night on the couch, for Anna Rohrbach they mean, above all, a lot of work.

Anna Rohrbach is a computer scientist. Together with her husband, Marcus, she is trying to teach computers something that might sound impossible at first: to watch videos and describe what is happening on screen. This is a trivial task for humans; at some point or other, we have probably all called out to the next room: "Honey, come quick! It's about to get really exciting!" When a gangster in a movie raises his weapon or the police chase a killer through dark alleyways, human viewers know exactly what's going on.

But a computer? First, a computer would have to be able to tell that a gun in a person's hand is a weapon and not a TV remote, that a hug has nothing to do with hand-to-hand combat, and that a fencing match isn't a matter of life and death. That in itself is a challenge. Then the moving images would need to be translated into comprehensible and grammatically correct natural language.

Anna and Marcus Rohrbach are experts in computer vision, which deals with automatic image recognition and analysis. Significant progress has been

Cooking on screen: Marcus Rohrbach set up a kitchen at the Max Planck Institute for Informatics and equipped it with video cameras. A computer program he developed is able to describe the cooking scenes being filmed here.

Software that learns: Marcus Rohrbach taught the computer program to recognize different activities being carried out in the kitchen by having assistants first describe the scenes. Here he is being assisted by doctoral student Siyu Tang.

made in this field over the past decade. Computers today can recognize faces in photographs and match them with different people. They can even correctly interpret pictures of landscapes. Reddish light, sails, horizontal lines? Sure thing: a sunset on the ocean. "But using clear words to correctly describe moving images in a movie scene is something else entirely," says Anna Rohrbach.

ONE APPLICATION IS IMAGE DESCRIPTIONS FOR THE BLIND

The scientist conducts research at the Max Planck Institute for Informatics in Saarbrücken. Marcus Rohrbach worked there too, before his recent switch to the University of California in Berkeley for a postdoc position, where he remains in close contact with his colleagues back in Saarbrücken. Yet the Rohrbachs aren't the only ones involved in this project. The idea originated from a collaboration between the Max Planck working group led by Bernt Schiele, in which Anna and Marcus Rohrbach work or worked, respectively, and the Department of Computational Linguistics at Saarland University, which is headed by Manfred Pinkal.

The researchers envision several applications for their project. In the future, computers could automatically generate and read out film descriptions for blind people. By today's standards, this is still a pretty costly and time-consuming process, because the voice-overs for movies need to be recorded by professional voice actors. A second possible application could be to automatically describe videos posted on online platforms. With the help of these short texts, Internet users could find relevant videos more quickly without first having to click through numerous clips.

A third application seems a bit more futuristic. If a computer is able to interpret movie scenes and describe them in natural language, then it can also comprehend events unfolding in the real world and render them in spoken words. That's why the Rohrbachs believe that, in just a few years, service robots or smartphone apps will be able to understand human actions and converse with humans using natural language. They could answer a user's question as to where he left his glasses, for example, or discuss what he should cook for dinner – after all, they observed which meals were served over the past few days.

Around five years ago, Marcus Rohrbach began teaching computers how to describe videos – a major goal that requires many small steps. "After all, you can't expect a software program to recognize the entire world with all its imaginable scenarios," the scientist explains. "That's why we decided to start out by limiting ourselves to one easily understandable scene – a kitchen, where we filmed people as they cooked." To this end, Marcus Rohrbach had a modern kitchen with a ceramic-glass cooktop and elegant cabinets specially set up at the Max Planck Institute.

Unlike a normal home kitchen, this one is fitted with several cameras that record what goes on in the room. The first step was to film volunteers as

Photo: MPI for Informatics

» The most important step: Marcus Rohrbach had to link the knowledge about movements and objects with activity descriptions – a complex process that is carried out in several stages.

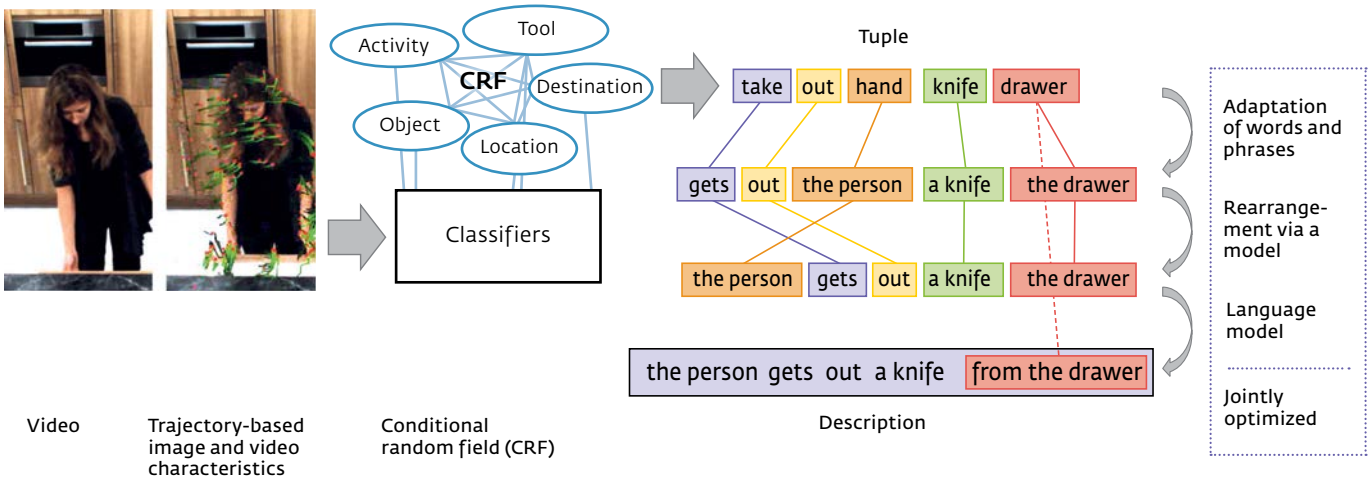
they performed different tasks – peeling an orange, cooking spaghetti or slicing a cucumber. Next, he gave his assistants the task of describing these film sequences using natural words – for example: "A man is standing in the kitchen and slicing a cucumber with the knife."

Since these descriptions are freely worded and have no fixed structure, the data then had to be annotated with comments that follow a fixed pattern. For example, the assistants noted down information pertaining to the following categories: object (such as a cucumber), activity (for instance peeling or slicing), tool (knife), location (countertop) and destination (salad bowl). "These categories are essential if you want to comprehensively describe an activity," Marcus Rohrbach explains, "because they contain the key elements of a sentence, such as the verb or object; for example: man – knife – slice – cucumber."

SOFTWARE TRACKS THE MOVEMENTS

Yet before a computer can describe objects, it first needs to learn what they look like. For this, Marcus Rohrbach used a software program that automatically learns the different parameters. The software is based on algorithms that are fed with a set of training data – in this case the video clips recorded in the kitchen. Step by step, the algorithm learns what an object looks like and can later recognize it accordingly. In the end, it computes probability values, such as, "This is 94 percent consistent with a banana."

Of course recognizing a video sequence also requires correctly identifying and interpreting movements. A rhythmically moving hand could be cutting, peeling a carrot or beating egg whites. The computer must be able to distinguish between these different actions. Marcus Rohrbach taught it such activities using tracking software. This software tracks the movement of individual pixels in a video image, essentially "freezing" the entire motion sequence. The researcher then fed this tracking data



Graphic: MPI for Informatics

The software developed by Marcus Rohrbach analyzes a video by first determining the image and video characteristics along the trajectories. Classifiers then identify objects, activities and tools as well as locations and destinations. Next, a model known as a conditional random field (CRF) creates a correlation between these parameters. This results in a tuple containing words and phrases that are first adapted to common wordings, for example by adding articles. The words and phrases are then rearranged before a language model adds any missing prepositions to form the final description.

Left Anna Rohrbach has collected around 200 DVDs in order to teach a software to describe the content of any given video.

Right In order to correctly describe a scene, a software must also be able to detect the subtext of the images. Otherwise it will mistake a hug for a wrestling match or the other way around, and it won't be able to distinguish between an Olympic fencing match and a historical duel.

into the algorithm as well, so that the computer learned to differentiate between cutting and peeling.

"These types of algorithms are known as classifiers," says Marcus Rohrbach. Depending on the probability value, they weigh different options to decide which action is being performed – for example cutting or stirring – or which object is involved – a cucumber or a banana. In order to do this, the classifier already has to take a range of characteristics, such as color, shape and size, into account when identifying the object.

A CONDITIONAL RANDOM FIELD PREDICTS THE ACTION

It's also important to model the interaction between various objects and activities. It's unlikely, for instance, that a person would peel a cucumber in a pot using a spoon; rather, you'd expect someone to stir zucchini in a pot using a wooden spoon, even though both scenarios might appear similar at first glance.

In order to predict which motion or activity is most likely being carried out, Marcus Rohrbach uses what is known as a conditional random field. This probabilistic model learns a correlation between the object, activity, tool and location. In other words, it predicts a group of categories, called a tuple; in this case, an object-activity-tool-location tuple. As with the other methods, the conditional random field model is also taught using training data.

The next step is the most important one. Marcus Rohrbach had to link this knowledge about movements and objects with activity descriptions – a complex process that is carried out in several stages. First, the classifier identifies the probability of individual elements. When a person puts an onion on the cutting board, the classifier will conclude that the following elements are the most probable: "hand", "put", "onion", "board", "countertop". The classifier excludes concepts that appear less probable, such as "spoon" or "pot". Next, the conditional random field computes which tuple best describes the given scenario – in this case, for instance: hand, put, onion, board.

"In order to then transform these tuples into natural language, we used an approach that translates texts, for example from English into German," says Marcus Rohrbach. As a first step, the software rearranges the concepts linked in the tuple to create a reasonable sequence, such as: "Hand put onion board."

Next, a language model adds any missing articles or prepositions to the words and phrases to form a semantically correct construct – in other words a sentence with a reasonable structure, such as: "The hand puts the onion on the board." In addition, it replaces certain terms with more commonplace wording that the language model is more familiar with – for instance "person" instead of "hand." Each computational step put together ultimately leads to the formation of a grammatically correct sentence, such as: "A person puts an onion on the board."

DETAILED DESCRIPTIONS VS. SHORT SUMMARIES

"The kitchen project was actually the topic of my Ph.D. thesis a while ago," Marcus Rohrbach explains. "This video description technique worked pretty well and correctly translated the scenes into natural language." Anna Rohrbach then expanded the model in such a way that it was able to describe scenes using different degrees of detail

Photo: Oliver Dietze



Photos: gruizza/iStockphoto, PaStar77/shutterstock, katatonia82/shutterstock (from left to right)

or abstraction – a feat that no other working group had accomplished before her. This method is thus capable of both describing the individual steps of an activity in detail, such as: "A woman takes spaghetti out of a cupboard, gets a pot out of the drawer and fills it with water," and summarizing the entire action in one concise sentence: "A woman cooks spaghetti."

Yet this first project had its limitations, says Marcus Rohrbach. After all, the video analysis system was limited to the kitchen setting. The whole system was also much too complex, in his opinion. The entire process of analyzing scenes, creating tuples, semantically correlating concepts and finally forming the finished sentence just seemed to take too long. "That's why we've set ourselves two new goals: we want to be able to analyze scenes in any given setting, and we want to reduce the whole process of turning a scene analysis into natural language output down to a single step."

This is where Anna Rohrbach's impressive film collection comes into play. She has analyzed 202 movies and 118,000 video clips to date. Each of these clips includes a natural language sentence description. She uses these data sets to train a special software tool: a long short-term memory (LSTM) network.

THREE CLASSIFIERS RECOGNIZE ONE SCENE

This tool is an artificial neural network that, like all software of this kind, mimics the functions of the human brain. Unlike other artificial neural networks, however, an LSTM remembers previously processed data over a longer period of time, which also allows it to process the input data more reliably when key signals (for example during scene recognition or speech) come in at irregular intervals.

Provided that such an LSTM is properly fed with training data, it can draw on its experience to independently decide which information is relevant and must be stored in the system, and which information can be deleted. This means the LSTM is capable of assessing the relevance of information. Today, LSTMs are often used for translating speech or recognizing handwriting.

An LSTM is the centerpiece of Anna Rohrbach's work. It links the visual information – the input – directly with the language generation, thus achieving the goal of reducing the video description process to a single step. The LSTM, too, uses probabilities. Its input is visual data, which in turn is supplied by classifiers. In order to fully recognize an entire scene, the scientist uses three different classifiers, which provide information about the following three aspects: the activity being performed, the objects in view, and the location in which the particular scene is taking place.

Anna Rohrbach also incorporates elements developed by other working groups, such as a classifier created by researchers at the Massachusetts Institute of Technology in the US. By feeding it a lot of data, the classifier was taught to recognize settings and environments – a kitchen, a bathroom, or a restaurant, for example. As usual, the classifiers supply probability values, which are then linked to form a probability vector – a cloud of probability values, if you will – before being fed into the LSTM.

The LSTM converts this visual information directly into natural language descriptions. "One of the strengths of this LSTM is that it can assess a sequence of words to predict which words are likely to follow," says Anna Rohrbach. It is very efficient at deciding which word must follow another

The neural network (LSTM) developed by Anna Rohrbach describes video sequences such as a dance scene more accurately than other computer programs, but not quite as well as a human just yet.

word, and at filtering out irrelevant data. The LSTM adds articles and prepositions, thus generating meaningful, natural language.

"It basically uses the same technique we humans do. We also remember which words we just said and formulate the next part of our sentence accordingly." Anna Rohrbach's LSTM has also developed what you could call a feeling for language. It no longer requires tuples that first string words together and then rearrange them step by step to form a complete sentence.

Ultimately, the LSTM uses probabilities to decide which word will come next. Apparently it does this very well: in a direct comparison, Anna Rohrbach's technique produced better results than other video description methods. Among other things, her LSTM was able to describe a scene with greater accuracy and more nuances than the other methods.

THE LSTM DELIVERS BETTER RESULTS THAN OTHER METHODS

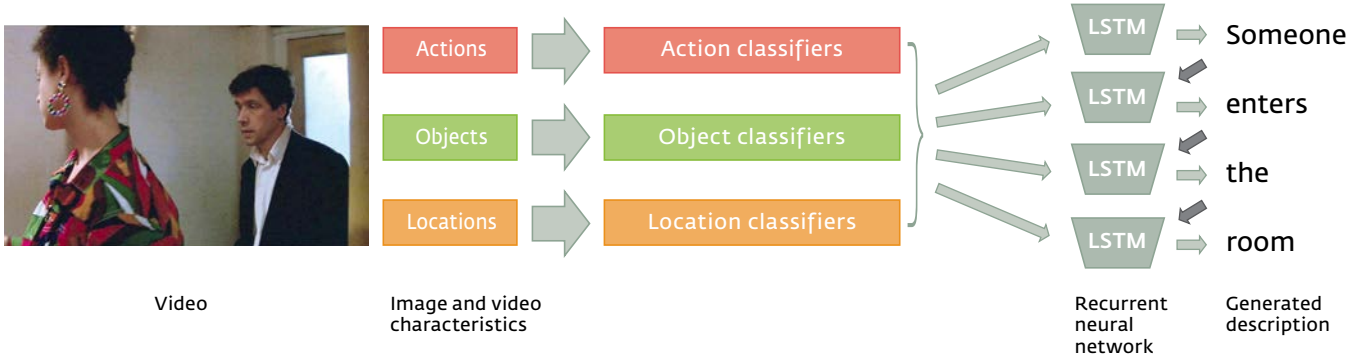
For example, a movie scene depicts a person leading a blonde woman onto the dance floor and then spinning her. Anna Rohrbach's LSTM described the scene as follows: "Someone is in a white dress, smiling with a smile and white hair." A different software offered a considerably less detailed description: "Someone glances at someone." The software developed by a third team against which Anna Rohrbach compared her LSTM even provided an unintentionally comic description of the two actors looking at each other: "Someone glances at someone. Someone glances at someone."

The comparison clearly shows that the LSTM analyzes the scene more accurately than other methods. At the same time, however, this example also exposes the weaknesses of Anna Rohrbach's system. After all, the LSTM didn't reveal that this scene took place in a ballroom. "It's true that this method isn't one hundred percent reliable yet. Grammar mistakes keep slipping in. And in some cases it doesn't correctly recognize scenes, especially when they are particularly complex," says the researcher.

One such example is a video sequence showing a young person in sports clothing running away. This scene was manually described for blind people as follows: "He runs up the steps of the stand and away." The LSTM interpreted: "Someone is running in the middle of the road."

This shows that the LSTM still has certain limitations, especially when it comes to recognizing abstract content. The LSTM wasn't able to make out that the young person is running away, and it also ignored the fact that he is running up a set of steps. "In other cases the system wasn't able to recognize that a person was fleeing from the police," says Anna Rohrbach.

Photo: Oliver Dietze



Anna Rohrbach's software learns to recognize actions, objects and locations depicted in a video using different classifiers, each of which is specific to one of these three categories. In a series of cycles, a recurrent neural network (LSTM) then uses these image characteristics to create a word-for-word description of the video.

"It's difficult to teach a computer to establish such thematic relationships between different pieces of content." Yet that is exactly what Anna Rohrbach has set out to achieve in the near future. She would also like to teach the computer to interpret actors' emotions. That would significantly improve the analysis method and bring video descriptions to a whole new level.

Rohrbach can't yet say exactly when her video description system will be ready to market. "But remarkable progress has been made in the field of image recognition over the past few years. So sometimes things can happen very quickly," she says. But she doesn't want to commit to anything just yet. The benefit for users would be substantial. Videos could be enhanced with text for the blind in no time at all. And Internet users could quickly skim through the content of online videos using either the concisely summarized description – "A woman cooks spaghetti" – or the extended, fully detailed text version. ◂

TO THE POINT

- **Over the past decade, significant progress has been made in the field of computer vision, which deals with automatic image recognition. For example, today's computers are able to recognize faces in photographs and attribute them to different people.**
- **Describing film scenes, on the other hand, is a much more complex process.**
- **Nevertheless, scientists hope to enable computers to automatically generate and read out video descriptions.**
- **To achieve this goal, researchers at the Max Planck Institute for Informatics are using a special software tool known as a long short-term memory (LSTM).**

GLOSSARY

Algorithm: A clear set of operations to be performed in order to solve a problem or class of problems. Algorithms consist of a finite number of individual steps and can be executed by being implemented in a computer program, for example.

Computer vision: The computer-aided approach of solving problems relating to the abilities of human vision. Possible applications include industrial production processes and traffic engineering.

Long short-term memory (LSTM): An artificial neural network that mimics the functions of the human brain and remembers previously processed data over a comparatively long period of time. When fed with training data, an LSTM can independently decide which information is relevant and must be stored in the system.

Graphic: MPI for Informatics