Das Gesicht als Interface zwischen Mensch und Maschine

WIE WIR ZUKÜNFTIG MIT ROBOTERN KOMMUNIZIEREN

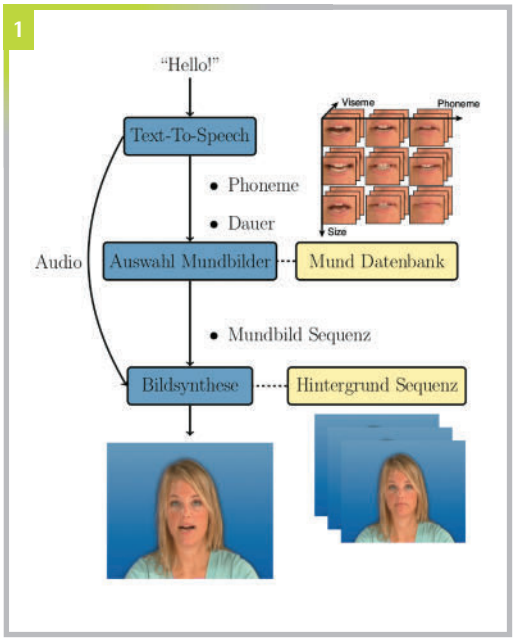
Der Roboter soll menschlicher werden: Dadurch soll einerseits die Akzeptanz der Maschinen verbessert, andererseits die Kommunikationsmöglichkeiten zwischen Mensch und Maschine gesteigert werden. Wissenschaftlerinnen und Wissenschaftler vom Institut für Informationsverarbeitung an der Fakultät für Elektrotechnik und Informatik forschen daher an der automatischen Erstellung eines hochaufgelösten, vielseitig einsetzbaren 3D-Gesichtsmodells.

Die natürlichste Form der menschlichen Kommunikation erfolgt nicht nur über die Sprache, sondern auch durch nonverbale Kommunikation, basierend auf Körpergesten und Gesichtsmimik. Trotzdem besteht die heutige Mensch-Maschine-Kommunikation auf menschlicher Seite hauptsächlich aus Interaktion via Texteingaben, Mausinteraktion und Fingergesten. Aktuelle Fortschritte sind Technologien zur Interaktion via Spracherkennung und Sprachausgabe, zum Beispiel Apples Siri, oder Interaktion via Körperbewegungen, die beispielsweise durch 3D-Sensoren wie Microsoft Kinect aufgezeichnet werden können.

Während Sprache und Körpergesten bereits in einigen Bereichen als Eingabe verwendet werden, hat sich bisher keine Technologie etabliert, die das menschliche Gesicht als Ein- oder Ausgabe verwendet. Dabei wäre das Gesicht, als Gesamtkonzept aus Mimik und Sprache, eine Möglichkeit, um künstlich erdachte Bedienkonzepte abzuschaffen und die Kommunikation zwischen Mensch und Maschine der zwischen Mensch und Mensch anzunähern.

Eine solche Form der Interaktion ist besonders wichtig für die zukünftige Generation an Robotern, die direkt im menschlichen Umfeld eingesetzt werden und autonom mit dem Menschen kommunizieren sollen. Zurzeit werden Roboter bereits als einfache Haushaltshilfe in unmittelbarer Nähe zum Menschen eingesetzt, beispielsweise zum Staubsaugen, Wischen oder Rasenmähen. In Zukunft werden Serviceroboter auch in Bereichen eingesetzt werden, die eine direkte soziale Interaktion mit dem Menschen und somit auch soziale Intelligenz des Roboters voraussetzen.

1



So genannte Personal Robots oder auch Social Robots werden bald Einzug in den modernen Haushalt finden, Aufgaben erledigen, überwachen und Prozesse nach den Wünschen ihrer Nutzer steuern. Bereits jetzt sind erste Formen solcher Roboter erhältlich oder vorbestellbar, beispielsweise Jibo und Personal Robot. Darüber hinaus ist der Einsatz sozialer Roboter in Bereichen der Pflege und Betreuung denkbar, um Personen mit Behinderung bei Aktivitäten des täglichen Lebens zu unterstützen und gleichzeitig Betreuungspersonal zu entlasten, zum Beispiel in der Altenpflege.

Ein wichtiger Schritt zu einer emotionalen und sozialen Intelligenz ist es, Robotern zu ermöglichen, die menschlichen Emotionen aus Sprache und Mimik zu verstehen und darauf in sinnvoller Weise zu reagieren. Ein synthetisches Gesicht, welches zum Ausdruck verschiedener Emotionen fähig ist, kann Roboter ein weiteres Stück menschlicher erscheinen lassen. Neben einer erweiterten Kommunikationsmöglichkeit steigert dies möglicherweise auch die Akzeptanz im menschlichen Umfeld sowie die Freude an der Nutzung solcher Roboter.

Das virtuelle Gesicht als Schnittstelle

An der Erstellung, Darstellung und Animation virtueller, also computergenerierter Gesichter wird bereits seit langem geforscht. Am Institut für Informationsverarbeitung beschäftigen wir uns mit virtuellen menschlichen Gesichtern, auch »Talking Heads« genannt. Unser Ziel ist es, eine personalisierte, realistische Gesichtsanimation vollkommen automatisch aus möglichst wenigen Eingangsdaten zu erzeugen.

Unser Talking Head Prototyp »Rebecca« wird als virtuelle Empfangsdame auf dem Bildschirm im Eingangsbereich des Instituts eingesetzt. Eine Demo-Version steht als Download auf unserer Website zur Verfügung. Rebecca kann beliebige Texte in audiovisueller Form präsentieren und hilft dem Benutzer durch die Menüs des Touchscreens zu navigieren.

Bei Rebecca handelt es sich um ein System, das auf einer Bilddatenbank basiert, wobei die Ausgabe durch eine Komposition von Bildern realisiert wird. Eine Übersicht zur Technik hinter dem System ist in *Abbildung 1* dargestellt. Der zu sprechende Text wird an einen Text-to-Speech (TTS) Server gesendet, der ein Tonsignal sowie zugehörige Lauteinheiten (Phoneme) und deren Dauer zur Verfügung stellt. Anschließend wird ein inverses Lippenlese-Problem gelöst, das heißt zu den jeweiligen Phonemen werden anhand eines Gütekriteriums plausible Mundbilder (Viseme) aus einer Mund-Datenbank ausgewählt und zu einer Mundbildsequenz zusammengesetzt. Diese wird dann auf das Gesicht der Hintergrundsequenz übertragen. Werden die Bild- und Tonsequenz synchron abgespielt, ergibt sich eine realistische Gesichtsanimation. Der große Vorteil solcher Systeme ist, dass jedes Einzelbild eine echte Aufnahme des Sprechers zeigt, so dass neue Sequenzen effizient und realistisch zusammengesetzt werden können.

Das vorgestellte System hat den Nachteil, dass nur Bildsequenzen aus Bildern erzeugt werden können, die bereits Teil der Datenbank sind. Aus diesem Grund bestimmen die Größe der Mundbild-Datenbank und die enthaltenen Laute und Emotionen die Qualität und Diversität der Synthese.

Als Alternative werden Gesichtsmodelle eingesetzt, die ein Gesicht durch kontinuierliche Parameter beschreiben und eine Synthese ermöglichen, die nicht auf eine feste Anzahl von Bildern beschränkt ist. Durch die Veränderung der Parameter kann ein neues Gesicht erzeugt werden. Gesichtsmodelle bieten auch die Möglichkeit für ein vorhandenes Gesicht die Parameter zu schätzen, die es beschreiben und zur Identifikation sowie zur Erkennung von Emotionen und Mimik verwendet werden können.

Ein Beispiel für ein kontinuierliches Modell ist Candide-3, wie in *Abbildung 2* dargestellt. Es besteht aus individuellen Shape-Parametern (zum Beispiel Augenabstand) und mimikbasierten Action-Parametern (zum Beispiel Augenbraue heben). Das Modell besteht aus 3D-Punkten, die durch Dreiecke verbunden sind und kann mit 14 Shape sowie 65 Action-Parametern individuell variiert werden.

Die Action-Parameter entsprechen dabei weitestgehend denen des Facial Action Coding System (FACS). Das FACS wurde 1978 von den Psychologen Ekman und Friesen definiert, um basierend auf den Gesichtsmuskeln die Gesichtsmimik anhand von kleinsten Bewegungseinheiten (Action Units) zu beschreiben. Der zweite Teil der Parameter entstammt den Facial Animation Parametern (FAP) des MPEG-4 Standards.

Ein Problem des Modells besteht darin, dass verschiedene Parameter Einfluss auf dieselben Punkte nehmen und ähnliche Veränderungen herbeiführen können. Werden diese Parameter zur selben Zeit geschätzt, ergeben sich unsinnige Werte und teilweise entstellte Gesichter.

Wir haben einen Algorithmus entwickelt, der korrelierte Gesichts-Parameter schätzen kann, dabei die Form eines Gesichts bewahrt und gleichzeitig zu einer deutlich höheren Genauigkeit führt. *Abbildung 2* zeigt, wie das Modell an ein vorhandenes Bild angepasst wird und durch eine Veränderung der Parameter ein neuer Gesichtsausdruck entstehen kann.

Abbildung 1
Unser Talking Head Rebecca ist als Demo-Download auf unserer Webseite verfügbar und wird im Eingangsbereich unseres Instituts auf einem Touchscreen verwendet.

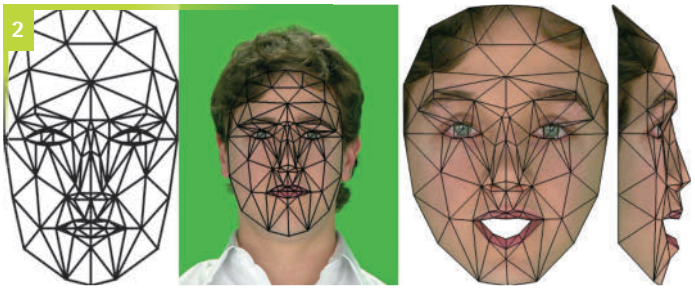
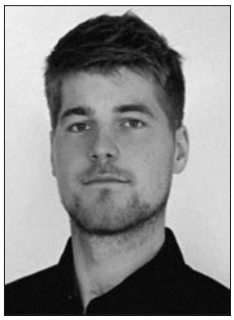


Abbildung 2
Von links nach rechts: Candide-3-Modell (von Ahlberg und anderen), Eingabebild einer Person mit adaptierten Modell, 3D-Modell mit einem geänderten Gesichtsausdruck.

Mit 113 Punkten bietet das Candide-3 nur eine grobe Vereinfachung eines Gesichts und eignet sich deshalb eher zur Analyse als zu einer realistisch wirkenden Synthese von Gesichtern.

Deutlich detailliertere, hochaufgelöste Modelle sind die in der 3D-Computergrafik häufig verwendeten Blendshape-Modelle. Diese modellieren ein Gesicht aus einer Linear-



Felix Kuhnke, M.Sc.
Jahrgang 1987, arbeitet seit 2015 als wissenschaftlicher Mitarbeiter am Institut für Informationsverarbeitung. Sein Forschungsschwerpunkt ist die Synthese visueller Sprache. Kontakt: kuhnke@tnt.uni-hannover.de

Stella Graßhof, M.Sc.
Jahrgang 1985, ist seit 2012 wissenschaftliche Mitarbeiterin am Institut für Informationsverarbeitung. Ihre Forschungsschwerpunkte sind 2D/3D-Gesichtsmodelle, Schätzung von Gesichtsparametern und Registrierung. Kontakt: grasshof@tnt.uni-hannover.de

Prof. Dr.-Ing. Jörn Ostermann
Jahrgang 1962, ist seit 2003 Leiter des Instituts für Informationsverarbeitung. Seine Forschungsinteressen sind Audio- und Videosignalverarbeitung, Computer Vision, 3D-Modellierung, Gesichtsanimation und Mensch-Maschine-Interaktion. Kontakt: ostermann@tnt.uni-hannover.de

kombination von Basisposen und bieten Anwenderinnen und Anwendern intuitive Interaktionen durch semantische Parameter. Die Modelle sind in der Regel kostenpflichtig und entstehen mit hohem manuellen Aufwand. Zudem besitzen diese Modelle nur Action-Parameter und sind somit auf eine Person festgelegt.

Die aktuelle Forschung beschäftigt sich daher mit der automatischen Erstellung eines hochaufgelösten, vielseitig einsetzbaren 3D-Gesichtsmodells, beispielsweise aus einer Datenbank von 3D-Gesichtern mit verschiedenen Personen und Gesichtsausdrücken.

Insbesondere die Anwendungen für visuelle Sprachsynthese sowie Transfer und Klassifikation von Gesichtsausdrücken sind dabei von Interesse.

Wo können Talking Heads eingesetzt werden?

Die Anwendungen für virtuelle Gesichter sind sehr vielfältig, zum Beispiel in Filmproduktionen und Videospielen. Im Folgenden werden zwei spezifische Anwendungsmöglichkeiten beschrieben.

Eine mögliche Anwendung für künstliche, sprechende Gesichter ist das visuelle Sprachverstehen. Schwerhörige und taube Personen kommunizieren oft durch Lippenlesen mit ihren Mitmenschen. Während ein traditionelles Telefonat für diese Menschen unmöglich ist, gibt es bereits Anwendungen, welche die Sprache in Text für den Gehörlosen übersetzen. Wir arbeiten daran, Sprache in eine realistische Gesichtsanimation zu übersetzen, die von so hoher Genauigkeit ist, dass sie das Lippenlesen für Schwerhörige und Gehörlose ermöglicht. Der große Vorteil dieser Technologie besteht darin, dass nur der Ton übertragen werden muss und die Bilder des sprechenden Kopfes auf einem Empfangsgerät erzeugt werden. Zusätzlich nutzen auch Normalhörende Lippenlesen, um in Situationen mit vielen Hintergrundgeräuschen, zum Beispiel am Bahnhof, Sprache besser zu verstehen. In diesen Bereichen kann eine zusätzliche visuelle Sprachausgabe die Sprachverständlichkeit auch für Normalhörende erhöhen.

Für den Bereich E-Care können personalisierte sprechende Köpfe eingesetzt werden, um die Kommunikation mit Menschen, die an Demenz erkrankt sind, zu vereinfachen. Vertraute Gesichter, zum Beispiel eines Angehörigen, können sie daran erinnern, genug zu trinken oder ihre Medikamente zu nehmen, und so die Pflegekräfte unterstützen. In solchen Fällen könnten entsprechende Personal Robots auch die Zeit verlängern, bevor pflegebedürftige Menschen in einem Pflege- oder Altersheim untergebracht werden müssen oder Angehörige bei der Pflege unterstützen. Daher ist auch ein »Talking Head für jedermann« ein Ziel unserer aktuellen Forschung. Aus einer Aufnahme einer Person soll vollautomatisch ein personalisiertes Modell erstellt werden, welches beliebig animiert werden kann, um zum Beispiel als sprechender Kopf auf dem Display eines Roboters zu erscheinen.

Weitere Entwicklungen im Bereich der virtuellen Gesichter aber auch die Zukunft der Mensch-Roboter-Interaktion bleiben spannend. Mit unserer Arbeit wollen wir Maschinen soziale Interaktion und somit auch soziale Intelligenz ermöglichen. Wann, ob und wo zukünftige Roboter mit uns zusammen lachen werden und dies auch über ein virtuelles Gesicht kommunizieren, bleibt offen.