Gesichtsanimation mit Image-based Rendering für Dialogsysteme

von

Dip.-Ing. Axel Weissenfeld

M.Sc. Kang Liu

Prof. Dr.-Ing. Jörn Ostermann

Institut für Informationsverarbeitung (TNT),

Leibniz Universität Hannover

Gesichtsanimation mit Image-based Rendering für Dialogsysteme

Inhaltsverzeichnis

Einleitung

Die heutige Mensch-Maschine-Kommunikation sieht bislang meist folgendermaßen aus: Der Mensch gibt eine Texteingabe an die Maschine, die dann entweder mit einer Text- oder einer Bildausgabe antwortet. In Zukunft kann die Maschine auch eine synthetische Sprachausgabe mit Gesichtsanimation zurückgeben, so dass die Gesichtsanimation als Teil einer modernen Mensch-Maschine Schnittstelle eingesetzt werden kann. Bei der Gesichtsanimation wird per Computer ein Gesichtsmodell zum Sprechen animiert und auf ein Display ausgegeben. Insbesondere interaktive Dialogsysteme, wie sie im Bereich des e-commerce und e-care (Customer Relation Management) zu finden sind, können eine Gesichtsanimation mit Sprachausgabe in ihre Web-Site integrieren und so die Kommunikation zwischen Mensch und Maschine verbessern. Dadurch steigt die Aufmerksamkeit und das Vertrauen des Nutzers zur Maschine .

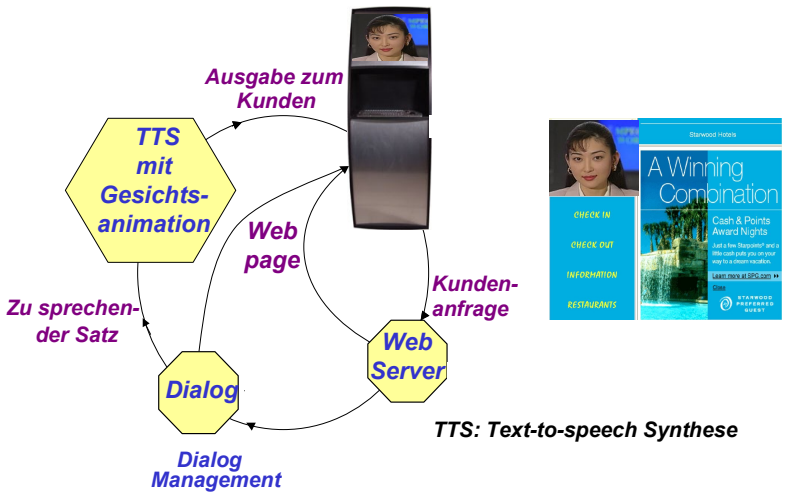


Abbildung 1: Hier ist die Architektur eines Dialogsystems im Bereich des e-commerce dargestellt, das über eine in die Web-Site integrierte Gesichtsanimation verfügt. Der Kunde kann eine Anfrage stellen, für die der Web Server eine Web-Site liefert. Außerdem generiert der Dialogmanager einen passenden Ausgabetext, der von einem TTS-Synthesizer in eine synthetische Sprachausgabe mit Gesichtsanimation umgewandelt wird.

In Abbildung 1 wird die Einbindung einer Gesichtsanimation in eine Web-Site dargestellt. Dem Kunden wird ein persönlicher, virtueller Einkaufsberater zur Verfügung gestellt, der ihn während des Einkaufs berät, aber auch unterhält. Subjektive Tests haben gezeigt, dass e-commerce Web-Sites, die eine Gesichtsanimation mit einer synthetischen Sprachausgabe in ihre Web-Site integrieren, eine höhere Kundenzufriedenheit erreichen . Die Bereiche des e-commerce und e-care zählen zu stark

wachsenden Industriesektoren, so dass in Zukunft die Gesichtsanimation verstärkt zum Einsatz kommen wird.

Heute reichen die Gesichtsanimationstechniken von der Animation eines 3D Gesichtsmodells bis Image-based Rendering. Ein 3D Gesichtsmodell, welches die Geometrie des Kopfes modelliert, besteht aus einem 3D-Polygonnetz, dessen Form durch Stützpunkte und Kanten definiert ist (siehe Abbildung 2). Zur Animation werden die Stützpunkte z.B. mit Hilfe von Muskelmodellen bewegt und das Polygonnetz texturiert. Bereits seit fast 30 Jahren wird im Bereich der Gesichtsanimation mit 3D Modellen geforscht. In werden Animationen erzeugt, indem verschiedene zuvor abgespeicherte Gesichtsausdrücke interpoliert werden. Mit steigender Rechenleistung konnten komplexere 3D Gesichts- und Steuerungsmodelle entwickelt und die Gesichtsanimation kontinuierlich verbessert werden. Jedoch erreichen die heutigen Animationssysteme basierend auf 3D Modellen noch keine photorealistische Animation. Unter einer photorealistischen Animation verstehen wir Animationen, die nicht von einem aufgenommenen Video zu unterscheiden sind.

Abbildung 2: Animation einer Sprecherin mit einem 3D Gesichtsmodell.

Im Jahre 1997 wurde erstmals in für die Gesichtsanimation auf das so genannte Image-based Rendering zurückgegriffen. Das Konzept des Image-based Rendering wurde bereits in den 60ern entwickelt, jedoch war der Rechen-und Speicheraufwand für praktische Anwendungen damals zu hoch. Beim Image-based Rendering werden mit Hilfe von vorher von einer Szene aufgenommenen Bildern neue Ansichten der Szene erzeugt. Dabei wird nur auf den Bilddaten gearbeitet, also neue aus bereits vorhandenen Bildern berechnet. Im Bereich der Gesichtsanimation werden die signifikanten Gesichtsmerkmale, wie zum Beispiel der Mund, aus einer aufgenommenen Videosequenz herauskopiert und in einer Datenbank gespeichert. Dieser Ansatz wurde in und vorgestellt und eine qualitativ gute Gesichtsanimation entwickelt. Das zur Zeit beste Verfahren für eine photorealistische Gesichtsanimation wird in beschrieben und dient als Grundlage für unser entwickeltes Gesichtsanimationssystems, welches im nächsten Kapitel beschrieben wird.

Gesichtsanimation mit Image-based Rendering

Die Gesichtsanimation besteht aus einer Gesichtsanalyse und Gesichtssynthese. Die Gesichtsanalyse legt eine Datenbank mit Mundbildern für jeden Sprecher an, der später animiert werden soll. Dazu werden Videosequenzen aufgenommen bei dem der Sprecher einen vorher definierten Korpus von einem Teleprompter abliest. Außerdem wird mit Hilfe eines 3D Scanners ein exaktes 3D Gesichtsmodell des Sprechers erzeugt. Die Gesichtssynthese muss in Echtzeit die eigentliche Animation erzeugen, damit nur eine kurze Latenzzeit auftritt.

Der erste Prozessschritt wird als Gesichtsanalyse bezeichnet. In diesem Schritt werden die Mundbilder aus den aufgenommenen Videosequenzen herauskopiert und in einer Datenbank abgelegt. Da ein Mensch den Kopf beim Sprechen bewegt, muss eine Schätzung der Position und Orientierung des Kopfes erfolgen. Die exakte Schätzung dieser Parameter ist essentiell für die spätere Animation, da der Mund sonst bei der Synthese zittert. Wenn die Position und Orientierung bekannt sind, können detektierte Gesichtsmerkmale normalisiert werden. Normalisieren bedeutet, dass die Gesichtsmerkmale so projiziert werden, dass die Variation der Orientierung und Lage des Kopfes kompensiert wird. Die normalisierten Mundbilder können dann in einer Datenbank abgelegt werden. Für jedes Mundbild werden die folgenden Parameter gespeichert: Die geometrischen Eigenschaften des Mundes, die Position und Orientierung des Kopfes und das entsprechende Phonem, welches dem Bild zugeordnet wird. Bei der Zuordnung der Phoneme zu den Bildern wird auf eine

Spracherkennungssoftware zurückgegriffen, dass jedem Bild der aufgenommenen Videosequenzen ein Phonem zuordnet.

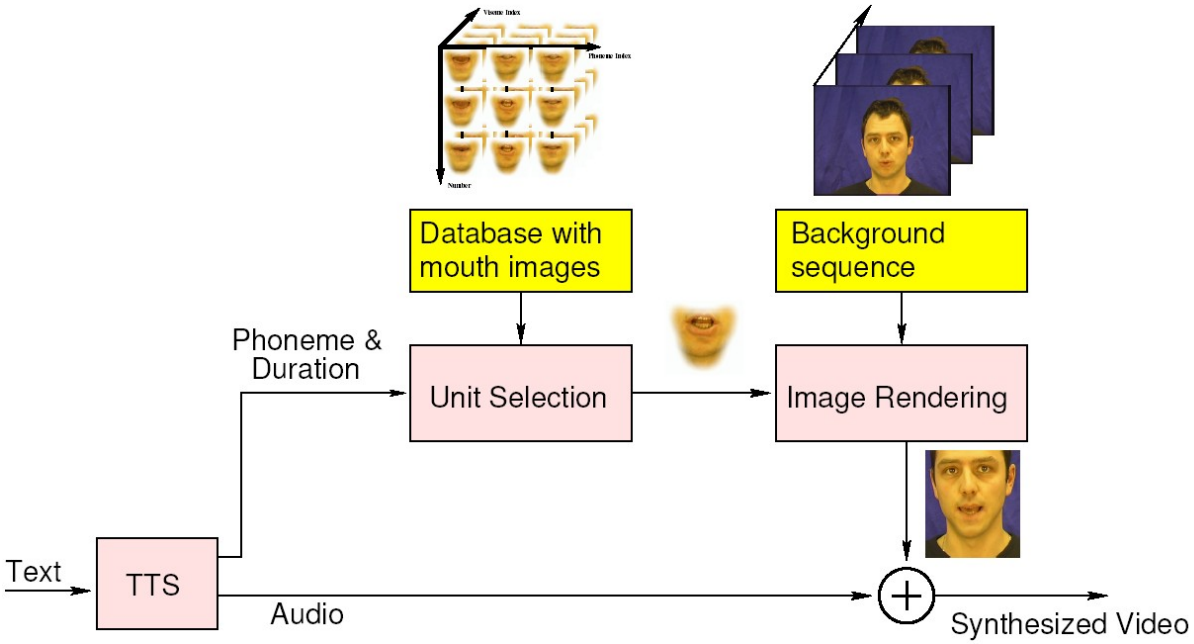


Abbildung 3: Das Blockschaltbild zeigt die Gesichtssynthese. Der TTS-Synthesizer sendet die Phoneme und deren Dauer zur Unit Selection. Die wählt aus einer Datenbank die passenden Mundbilder aus und diese werden dann in eine Hintergrundsequenz eingefügt.

Im zweiten Prozessschritt erfolgt die Gesichtssynthese (siehe Abbildung 3). Dort wird in einem ersten Schritt mit Hilfe eines TTS-Synthesizers ein Text zu einer synthetischen Sprachausgabe gewandelt und die einzelnen Phoneme mit deren Dauer bestimmt. Jedem Phonem wird der bestmögliche Mund aus der im ersten Prozessschritt angelegten Datenbank zugeordnet. Die Zuordnung der Mundbilder zu den Phonemen erfolgt unter Einhaltung eines Gütekriteriums. Diese Zuordnung ist nicht trivial, da man eine bestimmte Mundstellung nicht automatisch einem Phonem zuordnen kann. Der Mensch passt beim Sprechen seine Mundstellung schon einem zukünftigen Phonem an, bevor das Phonem selbst überhaupt ausgesprochen wird. Dieses Phänomen wird als Koartikulation bezeichnet. Dies wird bei der Animation gelöst, indem nicht nur das aktuelle Phonem, sondern auch vorherige und nachfolgende Phoneme bei der Auswahl des Mundes berücksichtigt werden. Die ausgewählten Mundbilder werden in eine Hintergrundsequenz eingefügt. Eine Hintergrundsequenz ist eine zuvor aufgenommene Sequenz des Sprechers, indem dieser typische, neutrale Kopfbewegungen ausführt. Beispiele für Animationen stehen unter http://www.tnt.uni-hannover.de/staff/aweissen/ zur Verfügung.

Zusammenfassung

Ein Gesichtsanimationssystem mittels Image-based Rendering wurde präsentiert. Die zwei Prozessschritte, Analyse und Synthese wurden kurz erläutert. Im Analyseschritt wird eine Datenbank mit den normalisierten Mundbildern erzeugt. Bei der Synthese werden Mundbilder so aus der Datenbank ausgewählt, dass die Bewegung des Mundes zur Sprachausgabe passt. Anschließend werden diese in eine Hintergrundsequenz eingefügt. Anwendungen dieser Gesichtsanimation liegen z.B. in webbasierten Dialogsystemen.

Referenzen

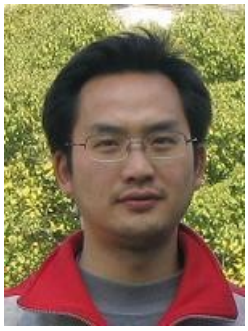
[1] J. Ostermann, D. Millen, “Talking heads and synthetic speech: An architecture for supporting electronic commerce,” Proc. ICME, pp. MA2.3, 2000.

[2] I. Pandzic, J. Ostermann, and D. Millen, "User Evaluation: Synthetic Talking faces for Interactive Services", accepted for publication in The Visual Computer, Special Issue on Real-time Virtual Worlds, 1999.

[3] J. Ostermann, “E-COGENT: An electronic convincing agent”, in MPEG-4 Facial Animation: The Standard, Implementation and Applications, Igor S. Pandzic (Editor), Robert Forchheimer (Editor), Wiley, Chichester, England, 2002, pp. 253-264.
[4] P. Ekman and W.V. Friesen: Manual for the facial action coding system, Consulting Psychologist Press, Inc. Palo Alto, CA, 1978.
[5] E. Cosatto, J. Ostermann, H. P. Graf, J. Schroeter, "Lifelike Talking Faces for Interactive Services," Invited Paper, Proc. of the IEEE, Special Issue on Human-Computer Multimodal Interface, Vol. 91, No. 9, pp. 1406-1429, Sept. 2003.
[6] T. Ezzat, G. Geiger, T. Poggio, "Trainable Videorealistic Speech Animation", Proc. ACM SIGGRAPH, pp. 388-397, 2002.
[7] C. Bregler, M. Covell, M. Slaney, "Video Rewrite: Driving Visual Speech with Audio," Proc. ACM SIGGRAPH, pp. 353-360, 1997.
[8] G. Kalberer, "Realistic Face Animation for Speech ", PhD Thesis, Swiss Federal Institute of Technology Zurich, 2003.
[9] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D. H. Salesin, "Synthesizing realistic facial expressions from photographs", In SIGGRAPH'98 Proceedings, pp. 75-84, 1998.
[10] K. Kaehler, J. Haber, H. Yamauchi, HP. Seidel, "Head Shop: Generating animated head models with anatomical structure", In Proc. 2002 ACM SIGGRAPH, Symposium on Computer Animation, pp. 55-63, 2002.

Verfasser-Portraits

Axel Weissenfeld erlangte 2003 das Diplom in Elektrotechnik an der Leibniz Universität Hannover mit Auszeichnung. Derzeit arbeitet er als Promotionsstipendiat der Deutschen Wirtschaft an seiner Promotion am Institut für Informationsverarbeitung an der Leibniz Universität Hannover. Seine Hauptarbeitsgebiete sind die Gesichtsanimation, Bildverarbeitung und Videocodierung. Für das beste Vordiplom im Bereich der Elektrotechnik bekam er 2000 eine Prämierung der Universität Hannover für hervorragende studentische Leistungen.



Kang Liu, Jahrgang 1977, beendete 2004 sein Studium der “Mechanical and Electronical Engineering” an der Zhejiang Universität, V.R.China und wird im gleichen Jahr wissenschaftlicher Mitarbeiter am Institut für Informations-vararbeitung, Universität Hannover. Seine Hauptarbeitsgebiete sind Gesichts-animation, Bildverarbeitung und Videocodierung.