Realistic face animation for speech

Diss. ETH No. 15277

Realistic Face Animation for Speech

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH

for the degree of
Doctor of Technical Sciences

presented by
Gregor A. Kalberer
Dipl. El. Ing.
born 4^{th} of April, 1972

Prof. Dr. Luc Van Gool, examiner
Prof. Dr. Thomas Vetter, co-examiner

2003

Abstract

This dissertation proposes an efficient work flow for realistic speech animation. It supports all steps of an entire animation pipeline, from the capture or design of three-dimensional (3D) head models up to the synthesis and editing stage of the performance. This pipeline is fully 3D, which yields high flexibility in the use of the animated character.

Realistic face animation for speech is still challenging, especially when we want to automate it to a large degree. Faces are the focus of attention for any audience, and the slightest deviation from normal faces and face dynamics is immediately noticed. This said, people would find it difficult to put their finger on what exactly it is that was wrong. We have to deal with subtle effects, that leave strong impressions. Hence, a convincing animation is a combination of very realistic head models with detailed and accurate facial dynamics, that are best adapted to the according head models.
At the input side of the proposed work flow, different options are open. A neutral head can be supplied either as a cylindrical scan of both range and texture data of a real person or as a face directly designed in a so called 'Face Space'. In terms of facial dynamics, real detailed 3D face motion, observed at video frame rate for thousands of points on the face of speaking actors, underpins the realism of the facial deformations. These deformations are given in a compact and intuitive representation via a 'Viseme Space' based on Independent Component Analysis. Performances amount to trajectories through this Viseme Space, whereas visemes are the visual counterparts of phonemes.
When asked to animate a face the pipeline replicates the visemes that it has learned, and adds the necessary co-articulation effects. Realism has been improved through a ground truth comparisons with real performance captured data.
Faces for which no 3D dynamics could be observed can be animated nonetheless. Their visemes are adapted automatically to their physiognomy by localizing the face in Face Space. Even as the system proposes an already convincing speech-based face animation, the computer cannot replace the creative component that a human expert brings in. The animator can thereafter still change the generated performance, that is given as a point of departure. Furthermore, the pipeline supports the animator with real captured emotional facial expressions and mechanical motions to increase the animation in terms of visual prosody.

Kurzfassung

Gegenstand dieser Dissertation ist es, eine effiziente Methodologie zur Generierung realistisch wirkender Gesichtsanimationen mit speziellem Fokus auf die visuelle Sprache zu finden. Alle wichtigen Schritte eines herkömmlichen Animationsablaufs wurden untersucht, behandelt und implementiert – von der Herstellung eines virtuellen Gesichtes bis hin zur Synthese einer überzeugenden Gesichtsanimation. Der vorgeschlagene Ansatz wurde für dreidimensionale (3D) Gesichter konzipiert, mit dem Ziel, höchstmögliche Flexibilität in Beleuchtung, Ansichtswinkel und Interaktion mit anderen 3D-Objekten zu bieten.

Trotz beträchtlicher Bemühungen während der letzten Dekaden, stellen uns realistisch wirkende Gesichtsanimationen auch heute noch vor grosse Probleme, besonders im Versuch, diese zu automatisieren. Das "Lesen" von Gesichtern ist sehr wichtig in unserem alltäglichen Leben und somit erstaunt es nicht, dass wir ausgesprochene Experten im Aufdecken kleinster Abweichungen gegenüber der Realität sind. Die kleinste Ungenauigkeit in einer Gesichtsanimation irritiert uns und lässt uns zweifeln. Diese Fähigkeit ist uns angeboren und seit frühester Kindheit an perfektioniert worden – ein tiefverwurzelter Instinkt, der nur sehr schwer getäuscht werden kann. Eine Gesichtsanimation muss also einerseits durch ein absolut realistisch aussehendes Gesicht und andererseits durch sehr detailierte organische Bewegungen überzeugen.

Ein neutrales Gesicht kann entweder in Form eines zylindrischen "Scans" unserem System übergeben oder direkt in einem system-internen "Face Space" hergestellt werden. Mit der Realitätstreue des vorliegenden Gesichtes, steigt die Erwartung an die Mimik proportional. Folglich werden sehr detailierte 3D Bewegungsdaten von sprechenden Gesichtern benötigt, um den hohen Erwartungen gerecht zu werden. Solche Daten wurden in unserem Falle mit einer Bildfrequenz von 24 Bildern pro Sekunde mittels eines speziellen 3D Rekonstruktionsverfahren aufgenommen, wobei tausende von 3D Punkten pro Bild extrahiert wurden. Diese enorme Datenmenge wurde anschliessend weiterverarbeitet und analysiert. Statistische Methoden (ICA) ermöglichten uns, sogenannte "Pseudo Muscles" zu isolieren. Sie spannen einen "Viseme Space" auf und repräsentieren die aufgenommenen Daten sowohl in einer sehr intuitiven als auch komprimierten Art und Weise. Die visuelle Sprache kann nun als Trajektorie durch diesen Viseme Space dargestellt werden. Die Viseme entsprechen visuell den Phonemen und werden durch Punkte representiert.

Soll nun ein Gesicht animiert werden, repliziert das System die Viseme, welche gelernt worden sind, und fügt die notwendigen Koartikulationseffekte hinzu. Die synthetisch generierten Trajektorien werden des weiteren mit aufgenommenen Trajektorien verglichen, um die resultierenden Animationkurven noch näher an die Realität zu approximieren.
Gesichter, für die keine dynamischen 3D Daten vorliegen, können nach unserem Ansatz trotzdem realistisch animiert werden. Ihre Viseme werden automatisch an ihre Physiognomie angepasst, indem das zu animierende Gesicht in den Face Space projeziert wird und die entsprechenden Viseme anhand der jeweiligen Position kombiniert und berechnet werden.
Obwohl die resultierenden Gesichtsanimationen bereits sehr überzeugend wirken, können Computer noch lange nicht die Kreativität eines erfahrenen Animators ersetzen. Deshalb bleibt es während jedem einzelnen Schritt dem Anwender überlassen, die vom Computer vorgeschlagene Schritte zu verfeinern.