Fortschritt-Berichte VDI

VDI

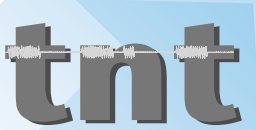
Reihe 10

Informatik/
Kommunikation

Stella Graßhof, M.Sc.,
Kopenhagen

Nr. 868

Expressive Personalized 3D Face Models from 3D Face Scans



Institut für Informationsverarbeitung
www.tnt.uni-hannover.de

Expressive Personalized 3D Face Models from 3D Face Scans

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor-Ingenieurin

(abgekürzt: Dr.-Ing.)
genehmigte

Dissertation

von

Stella Graßhof, M. Sc.
geboren am 20. August 1985 in Hannover.

2019

Hauptreferent:	Prof. Dr.-Ing. Ostermann
Korreferent:	Prof. Dr.-Ing. Rohs
Vorsitzender:	Prof. Dr.-Ing. Rosenhahn
Tag der Promotion:	08.11.2019

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Stella Graßhof, M.Sc.,
Kopenhagen

Nr. 868

Expressive Personalized 3D Face Models from 3D Face Scans

Institut für Informationsverarbeitung
www.tnt.uni-hannover.de

Graßhof, Stella
Expressive Personalized 3D Face Models from 3D Face Scans
Fortschr.-Ber. VDI Reihe 10 Nr. 868. Düsseldorf: VDI Verlag 2020.
216 Seiten, 57 Bilder, 6 Tabellen.
ISBN 978-3-18-386810-0, ISSN 0178-9627,
€ 76,00/VDI-Mitgliederpreis € 68,40.

Keywords: 3D face scans – nonrigid registration – correspondence estimation – expression intensity – tensor – factorization – statistical models – expression transfer – 3D reconstruction

Für die Dokumentation: 3D-Gesichts-Scan – nicht-rigide Registrierung – Korrespondenzschätzung – Intensität von Gesichtsausdrücken – Tensor – statistische Modelle – Transfer von Gesichtsausdrücken – 3D-Rekonstruktion

In this work, different methods are presented to create 3D face models from databases of 3D face scans. The challenge in this endeavour is to balance the limited training data with the high demands of various applications.
The 3D scans stem from various persons showing different expressions, with varying number of points per 3D scan and different numbers of scans per person. This data of posed facial expressions revealed substructures, which are utilised to improve the proposed model. In the process of creating and using the models, for each specific application objective quality criteria are carefully designed tailored to the task to quantify the quality.
In total four face models built from three databases are compared based on: 3D face synthesis, 3D approximation, person and expression transfer, and 3D reconstruction from 2D.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek
(German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

Als Manuskript gedruckt. Printed in Germany.
ISSN 0178-9627
ISBN 978-3-18-386810-0

Acknowledgment

First and foremost, I thank my doctoral advisor Prof. Dr.-Ing. Ostermann for giving me the opportunity to start and finish this work under his supervision at the Institut für Informationsverarbeitung (TNT), Leibniz University of Hannover, Germany. I appreciate the time dedicated for discussions, and guidance. I thank Prof. Dr.-Ing Rohs, who took the time to be my second supervisor and who contributed valuable hints to improve the final thesis. At the TNT I valued the discussion which I had with Prof. Dr.-Ing. Bodo Rosenhahn, whom I also thank for being the chair of my defense committee.

At the TNT several people helped me with software, hardware, and bureaucracy, which made my life easier and enabled me to focus on more relevant things. For lifting the burden of dealing with formalities, I thank Martin Pahl, and Thomas Wehberg, as well as the secretaries Doris Jasper-Göring, Hilke Brodersen, Pia Bank, and Melanie Huch. Many thanks go to Matthias Schuh for his hardware support, Marco Munderloh, Martin Pahl, Arne Ehlers, and Holger Meuel for resolving many software- and operating system-related issues, and who sometimes provided support during some unconventional hours.

During my time at the TNT I had the great honor to learn a lot from different fields from the nicest colleagues I could have wishes for. Thank you all for making this a great place to work. Particularly I thank my former office mates and discussion partners Matthias Reso, Felix Kuhnke, Benjamin Spitschan, and Hanno Ackermann.

Finally, very special thanks go to my family, especially my parents, Frank and Veronika Graßhof, who supported me and encouraged me to pursuit a career path, which feels right for me. Last but not least, I am grateful for my husband, Patrick Fließ, who believed in me, when I did not, and supported me with his patience and understanding.

Contents

Abstract

The creation of versatile 3D face models from limited training data has been a long-standing goal in facial animation. These models need to fully represent each individual face shape, including changes of facial expressions without the loss of individual facial features.

This difficulty is especially well-known in the movie industry, where even nowadays extensive manual work is necessary to achieve a natural representation of a human face with convincing expressive performance. This process is already challenging if sufficient high-quality 3D material of one person is available, but is considerably more difficult in the case of low-quality input data caused by limited hardware. In this work, different methods are presented to create 3D face models from databases of 3D face scans. The databases contain scans of various persons showing different expressions, a variety of points per 3D scan and different numbers of scans per person. Throughout this work objective quality criteria are carefully designed to quantify the quality requirements of each specific application.

In the first part of this work a preprocessing pipeline is presented, followed by a procedure to achieve dense meaningful correspondences between the 3D face scans. Then, based on the assumption of a shared motion pattern, a temporal alignment is estimated, which provides the same number of scans per person, such that facial motions are performed in synchrony. In this process, a robust descriptor for expression intensity is proposed, for which additional applications are presented, e.g. person-specific emotion cluster unveiling a variance in performance for each emotion between persons.

Since the resulting 3D faces are in full dense point-wise correspondence and their temporal facial movements are synchronized, they are aligned in space and time. Therefore, the processed 3D face scans can be arranged into a single data structure representing a 3D cube with axes corresponding to the number of 3D points, subjects and expressions, respectively. This data structure is referred to as *tensor* and outperforms the separation of different modes of individual shape and expression compared to traditional approaches based on 2D data structures, i.e. matrices. A 3D face model

is created from the data tensor by factorization into different modes. In contrast to former methods, the structures of the expression subspace are employed to derive reasonable constraints. In the expression subspace, it is observed that the six basic emotions (anger, disgust, fear, happiness, sadness, surprise) performed in different strengths, each form linear trajectories within the subspace. These six lines, each corresponding to one emotion, converge at a point which defines the natural origin of all expressions. It appears that this specific expression is not part of the database and that it differs from the neutral expression. Due to the fact that the database is based on posed instead of spontaneously performed expressions, the expression labeled as neutral differs from the fully relaxed face which would represent the expected case. Therefore the newly discovered origin is referred to as *apathetic*, which corresponds to an expression with fully relaxed facial muscles. It can be used for various applications: (1) to neutralize faces and replace the face with the original label neutral in the database, thereby improving the quality of the originally posed data without the need for new recordings, (2) to synthesize more convincing facial animations with an improved separation of distinct emotions, and (3) to adapt the statistical face model to render it more compact and robust, thereby enabling to perform stable expression and person transfer.

In this work four different face tensor models based on three databases are presented and compared for different applications: (1) 3D face synthesis, (2) 3D approximation, person and expression transfer, and (3) 3D reconstruction from 2D. The experiments show that dense 3D face reconstructions from sparse 2D landmarks based on the proposed models outperform those of the two state-of-the-art methods, although they employ more information from the original image.

Keywords: 3D face scans, 3D faces, nonrigid registration, correspondence estimation, expression intensity, tensor, factorization, statistical models, expression transfer, 3D reconstruction

Kurzfassung

Schon lange arbeiten Menschen daran aus begrenzten Trainingsdaten vielseitig einsetzbare 3D-Gesichtsmodelle zu erzeugen. Diese sollen einerseits das Gesicht gut repräsentieren und andererseits glaubhafte Änderungen des Gesichtsausdrucks ermöglichen. Insbesondere in der Filmindustrie ist das Problem bekannt ein überzeugendes Ergebnis eines menschlichen Gesichts mit natürlichen Gesichtsausdrücken zu erzeugen und erfordert noch heute viel manuelle Arbeit. Trotz der Verfügbarkeit hochqualitativer Daten ist dies weiterhin eine Herausforderung, insbesondere dort, wo limitierte Hardware weniger gute Ergebnisse liefert. In dieser Arbeit werden Ansätze präsentiert, um verschiedene 3D-Gesichtsmodelle zu erzeugen. Diese basieren auf Datenbanken mit 3D-Scans von Gesichtern, die jeweils verschiedene Personen und Gesichtsausdrücke enthalten, sich jedoch in der Anzahl der Punkte und Scans pro Person unterscheiden. Zudem werden in jedem Teil dieser Arbeit objektive Qualitätskriterien definiert, die jeweils Eigenschaften speziell für die jeweilige Anwendungen quantifizieren.

Im ersten Teil dieser Arbeit wird eine zielgerichtete Vorverarbeitung der Daten präsentiert, gefolgt von einem Ansatz, um sinnvolle Korrespondenzen zwischen den 3D Gesichts-Scans zu schätzen. Unter der Annahme, dass es ein gemeinsames Bewegungsmuster in mehreren Aufnahmen von Gesichtsausdrücken gibt, wird eine zeitliche Ausrichtung geschätzt, um dieselbe Anzahl von Scans pro Person zu erhalten, so dass Gesichtsbewegungen synchron erfolgen. Dabei werden ein Deskriptor für die Intensität des Gesichtsausdrucks definiert und weitere Anwendungen präsentiert.

Die verarbeiteten 3D-Gesichts-Scans sind nun in Zeit und Raum sinnvoll geordnet. Daher können diese in eine Datenstruktur sortiert werden, die einem Würfel entspricht, bei dem die drei Dimensionen folgende Informationen enthalten: Anzahl der 3D-Punkte, der Identitäten und der Gesichtsausdrücke. Diese Datenstruktur wird als Tensor bezeichnet und erleichtert die Trennung von individueller Gesichtsform und Gesichtsausdruck im Vergleich zu traditionellen Methoden, welche die Daten in eine 2D-Datenstruktur, d.h. Matrizen, einordnen. Basierend auf dieser Datenstruktur wird ein Gesichts-

modell mit einem Faktorisierungssatz erstellt. Anders als vorangegangenen Arbeiten, werden hier die gefundenen Strukturen in den Unterräumen verwendet um sinnvolle Nebenbedingungen zu definieren. In einem Unterraum finden sich Strukturen, in denen die sechs prototypischen Emotionen (Ärger, Ekel, Angst, Glück, Trauer, Überraschung), die in unterschiedlicher Stärke ausgeführt wurden, jeweils eine Gerade in dem Unterraum bilden. Diese sechs Geraden, jeweils zugehörig zu einer Emotion, treffen sich in einem gemeinsamen Punkt, welcher dem natürlichen Ursprung aller Gesichtsausdrücke entspricht. Es stellt sich heraus, dass dieser Gesichtsausdruck nicht Teil der Datenbank ist und sich von dem als neutral gekennzeichneten Gesichtsausdruck unterscheidet. Basierend darauf, dass die Datenbank aus Aufnahmen von gestellten und nicht aus spontan ausgeführten Gesichtsausdrücken besteht, folgt, dass Gesichter, die als neutral gekennzeichnet sind, individuelle Merkmale enthalten, die nicht immer dem erwarteten neutralen Gesichtsausdruck entsprechen, nämlich einem entspannten Gesichtsausdruck. Daher wird der gefundene neue Ursprung als *apathischer* Gesichtsausdruck bezeichnet, da er einem Ausdruck entspricht, bei dem alle Gesichtsmuskeln vollständig entspannt sind. Dieser kann für verschiedene Anwendungen genutzt werden: (1) Nachträgliche *Neutralisierung* des Gesichtsausdrucks, um die originalen Daten mit dem Label *neutral* zu ersetzen, wobei die Qualität der Daten verbessert werden kann ohne neue Aufnahmen zu benötigen. (2) Synthese überzeugender Gesichtsanimationen, welche die Vermischung verschiedener Emotionen verhindert. (3) Darüber hinaus wird demonstriert wie statistische Gesichtsmodelle robuster gemacht werden können, so dass der Austausch von Gesichtsausdrücken und Personen stabilisiert wird.

In dieser Arbeit werden vier verschiedene tensorbasierte Gesichtsmodelle, erstellt aus drei Datenbanken, vorgestellt und anhand verschiedener Anwendungen verglichen: (1) Synthese von 3D-Gesichtern, (2) 3D-Approximation, Transfer von Gesichtsausdrücken und Identität, und (3) 3D-Rekonstruktion aus 2D-Input. Es wird gezeigt, dass die präsentierten 3D Rekonstruktionen basierend auf wenigen 2D-Landmarken, die durch die vorgestellten Modelle erzeugt wurden, bessere Ergebnisse liefern als zwei State-of-the-Art-Methoden, obwohl diese mehr Informationen aus den Bildern verwenden.

Stichworte: 3D-Gesichts-Scan, 3D-Gesichter, nicht-rigide Registrierung, Korrespondenzschätzung, Intensität von Gesichtsausdrücken, Tensor, statistische Modelle, Transfer von Gesichtsausdrücken, 3D-Rekonstruktion

Abbreviations and Nomenclature

Abbreviations

3DMM 3D Morphable Model

AU Facial Action Unit

CPD Coherent Point Drift

DTW Dynamic Time Warping

ECPD Extended Coherent Point Drift

EM Expectation Maximization

FACS Facial Action Coding System

ffp facial feature point

FAP Facial Animation Parameter

FAU Facial Action Unit

GCTW Generalized Canonical Time Warping

GMM Gaussian Mixture Model

HOSVD High-Order Singular Value Decomposition

ICA Independent Component Analysis

ICP Iterative Closest Point

LFAU Lower Facial Action Unit

MAP Maximum A Posteriori

ML Maximum Likelihood

PCA	Principal Component Analysis
pdf	probability density function
SVD	Singular Value Decomposition
UFAU	Upper Facial Action Unit

Nomenclature

$\mathbb{C}$	complex numbers
$\mathbb{R}$	real numbers
$\mathbb{N}$	natural numbers
$\mathbb{E}(y)$	expectation value of random variable y
s	lower case italic letters define scalar values: $s \in \mathbb{R}$.
$\boldsymbol{v}$	lower case bold letters define column vectors: $\boldsymbol{v} \in \mathbb{R}^{N \times 1}$
$\boldsymbol{M}$	upper case bold letters define matrices: $\boldsymbol{M} \in \mathbb{R}^{M \times N}$
$\mathcal{T}$	upper case slanted letters define sets or tensors, e.g. a 3D tensor $\mathcal{T} \in \mathbb{R}^{L \times M \times N}$
v_i	ith element of vector $\boldsymbol{v}$
m_{ij}	element of ith row and jth column of matrix $\boldsymbol{M}$
$\boldsymbol{M}(i,:)$	ith row of matrix $\boldsymbol{M}$
$\boldsymbol{M}(:,j)$	jth column of matrix $\boldsymbol{M}$
$\boldsymbol{M}(:)$	concatenate columns of matrix $\boldsymbol{M}$ to vector
$\boldsymbol{v}^{\mathrm{T}}$, $\boldsymbol{M}^{\mathrm{T}}$	transposed vector $\boldsymbol{v}$ and matrix $\boldsymbol{M}$
$\boldsymbol{M}^{-1}$	inverse matrix
$\mathbf{1}_n$	vector of n ones: $\mathbf{1}_n \in \mathbb{R}^n$
$\boldsymbol{I}_n$	unity matrix $\boldsymbol{I}_n \in \mathbb{R}^{n \times n}$
$\lvert a \rvert$	absolute value of scalar value
$\lvert \boldsymbol{a} \rvert$	length of vector $\boldsymbol{a}$

$\lvert\boldsymbol{A}\rvert$	determinant of matrix $\boldsymbol{A}$
$\lVert\boldsymbol{v}\rVert_0$	0-norm of vector $\boldsymbol{v}$
$\lVert\boldsymbol{v}\rVert_1$	1-norm of vector $\boldsymbol{v}$
$\lVert\boldsymbol{v}\rVert_2$	Euclidean norm of vector $\boldsymbol{v}$
$\lVert\boldsymbol{M}\rVert_F$	Frobenius norm of matrix $\boldsymbol{M}$
$\mathbb{1}(A)$	the indicator function evaluates to either 0 if A is false and 1 if A is true
δ_{ij}	Kronecker delta is 0 if $i \neq j$ or 1 if $i = j$
$\otimes$	Kronecker product of vectors or matrices
$\mathcal{N}(\mu,\sigma^2)$	Gaussian distribution with expectation value μ and variance σ^2
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	multivariate Gaussian distribution with vector-valued expectation value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{T} \times_k \boldsymbol{M}$	mode-k tensor product between tensor $\mathcal{T}$ and matrix $\boldsymbol{M}$
$\ln(\cdot)$	logarithm to base e
$\mathrm{diag}(\boldsymbol{v})$	returns diagonal matrix with input vector on the diagonal
$\mathrm{diag}(\boldsymbol{M})$	extracts diagonal from matrix $\boldsymbol{M}$ as vector
$\mathrm{tr}(\boldsymbol{M})$	trace of matrix $\boldsymbol{M}$