MIRROR-BASED MULTI-VIEW ANALYSIS OF FACIAL MOTIONS

Jürgen Rurainsky and Peter Eisert

Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute
Einsteinufer 37, 10587 Berlin, Germany

ABSTRACT

We present our system for the capturing and analysis of 3D facial motion. A high speed camera is used as capture unit in combination with two surface mirrors. The mirrors provide two additional virtual views of the face without the need of multiple cameras and to avoid synchronization problems. We use this system to capture the motion of a person's face while speaking. Investigations of these facial motions are presented and rigid and non-rigid motion are analyzed. In order to extract only facial deformation independent from head pose, we use a new and simple approach for separating rigid and non-rigid motion named Weight-Compensated Motion Estimation (WCME). This approach weights the data points according to their influence to the desired motion model. We also present first results of our model-based facial deformation analysis. Such results can be used for facial animations in order to achieve a higher degree of quality.

Index Terms— 3D modeling & synthesis, parametric models for motion estimation, mirror, multiview, facial deformation

1. INTRODUCTION

Different capture systems for the analysis of facial motions have been presented in recent years. Although single or multi camera approaches have been addressed with different configurations, mirrors are rarely used. One reason for this could be the resolution of the capture unit, which is shared with all virtual views.

Since we target for the analysis of the dynamic behavior of facial motion, the sampling rate, in which the motion states are recorded, is an important issue. Important transitions from one state to another get lost if only a video frame rate of 25 fps is used and these details are not available for the natural animation of 3D models.

High-end motion capture systems, as used for movie productions, can realistically animate another object, a person, or a creature by mapping an actor's motion to it [1]. Rather than only animating faces with the motion information, facial motion and specific facial states are also analyzed for medical purposes, treatment, and diagnosis [2, 3]. In this case, the resolution of the analyzed facial motion is mostly limited to the anatomically interesting points and is focused to facial expressions rather than facial motion caused by speech.

Although different approaches for the specification of static expressions are available like the Facial Action Coding System or the MPEG-4 Facial Animation Parameters FAPs, much less has been reported about the dynamic modeling of these motions. In [4], an dynamic extension to FACS system is presented. In [5, 6] results of 3D speech movement analysis by using facial deformation states are given and used for animation and tracking purposes. These results show promising gains and lead to higher degrees of acceptance in facial animation.

We present our capture system based on a high speed camera and two surface mirror. These three views are used to reconstruct a 3D model sequence of a talking person's face. In order to analyze only local facial deformation and not the global head movements, we introduce a new and simple approach for the separation of rigid-body and non–rigid motion named Weight Compensated Motion Estimation (WCME). Additionally, initial results for the analysis of non-rigid deformations are provided.

2. CAPTURE SYSTEM

Mirror constructions can be used to simultaneously capture more than one view using a single capture device. In [7], a mirror construction is described to capture the movements of the lips from two views. A double mirror construction together with two cameras is described in [6] and used to determine speech movements from a total of four views, where the real camera views are almost similar. We have constructed a system with two mirrors which is shown in **Fig. 1**. Surface mirrors are used, in order to avoid multiple reflections of the recorded object. A high speed camera running at 200 fps and a resolution of 1536 x 1024 pixels is used for capturing. High speed capturing, even performed at only 200 fps, requires an

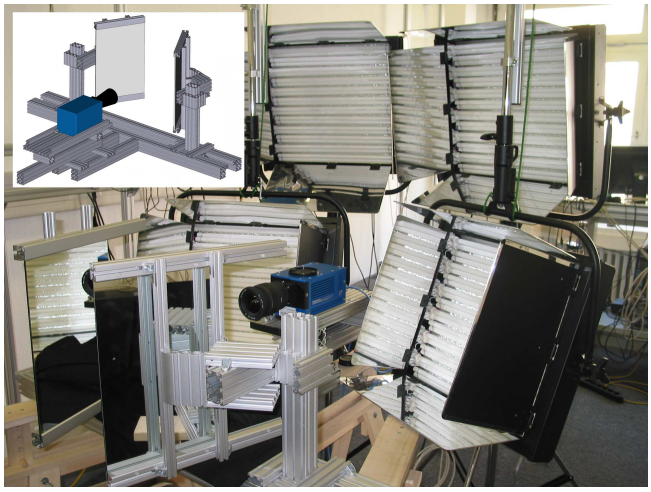

Fig. 1. High speed camera and mirror construction with two surface mirrors. Top Left Corner: Schematic scheme of the construction.

appropriate amount of scene illumination. Four flat lights as shown in **Fig. 1** are used to illuminate the scene uniformly. The mirrors are located almost symmetrically to the z-axis and adjusted to an angle of $\pm 31^{\circ}$. Other mirror DOFs are adjusted to achieve a good balance between all three views in the camera frame.

2.1. Captured Material

We use our mirror construction from **Fig. 1** to capture a talking person's face which is shown in **Fig. 2**. Due to the selected frame rate of 200 fps and the amount of memory located on the capture device, each sequence consists of 2048 frames.
The test person's face was covered with around 150 markers made out of dark green fabric tape with a physical dimension of about 2 x 2 mm. Two constraints are considered during placement of the markers onto the test person's face: uniform distribution and good visibility during facial deformation. The markers on the upper and lower lip were placed in alternating rows, in order to avoid interference problems during tracking.
The person was asked to look directly into the camera and to return to this view after performing the requested facial deformation. During the capture process, only small rotation angles and translation were permitted.

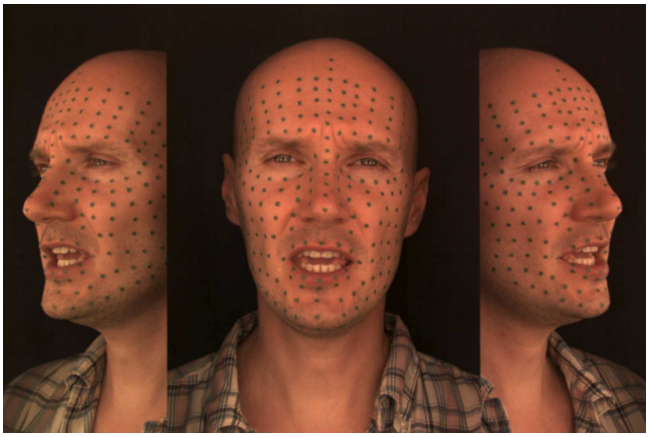

Fig. 2. Captured Material: One frame of the captured sequence with a resolution of 1536 x 1024 pixels at 200 fps. Multiple view recording using two virtual cameras turned $\pm 31^{\circ}$ to z-axis.

2.2. Calibration

Since the 3D position of all markers needs to be computed for each frame, a calibration of the entire setup is required. This includes the determination of the intrinsic camera parameters that specify the projection, but also the position and orientation of the two virtual cameras. A model-based calibration techniques [8], that exploits the entire image information, is used to accuratly estimate the camera parameters. For this particular setup, the calibration framework needs to consider that the two virtual views are flipped horizontally and have exactly the same intrinsic parameters as the real camera. Therefore, a joint estimation for a single set of intrinsic parameters (radial lens distortion, f_x, and f_y) and two sets of extrinsic parameters is performed and used in the following to determine the exact 3D location of the markers.

3. RIGID-BODY MOTION ESTIMATION

The extraction of non-rigid deformation requires a differentiation between rigid and non-rigid movements. Rigid body motion can be described by 6 DOFs (rotation and translation for all axes) of the associated 3D model and all other changes are regarded as deformation.

Rigid-body motion and deformations are very successfully determined by several different approaches. In [9, 10, 11], methods for motion estimation from a single view using optical flow are described. A neural network was formed to estimate the rigid-body motion in [12] using multiple views. In [13] a simulated annealing approach was introduced to determine the desired motion parameters. The classification of the available 3D model vertices into a rigid and a non-rigid class is described in [14].
We present a new and simple approach named Weight Compensated Motion Estimation (WCME) to estimate the rigid-body motion parameters in the presence of non-rigid deformation. This approach is applied to existing 3D models and continuously separates the vertices into rigid and non-rigid motion.

3.1. Model Reconstruction

After correcting the white balance and radial lens distortions for all frames of the sequence, the markers are searched in the images. In a first analysis step, the 2D marker positions are tracked in all views and their 3D position is computed using calibration data. Starting from the first frame, the relation between all marker positions in an image is used to support the tracking process of the markers over the entire sequence, e.g. motion vectors for hidden markers were estimated by considering motion vectors of connected markers. A triangle mesh describes the topology of the markers.
The correspondences between the middle and left view and the middle and right view, respectively, are defined as well. These correspondences are used for each frame to reconstruct two 3D models by triangulation using the middle view as reference. The resulting two 3D models were combined at the common vertices. This results in a sequence of topologically identical 3D models. Each 3D model consists of 127 vertices, 175 triangles and a per vertex labeling of visibility.

3.2. Motion Model

Because of the very small expected relative rotations and translations between successive frames, a linearized version of the rotation matrix is used to determine the rigid body motion. More details of this linearized version are given in [10]. The use of linearized rotation parameters leads to computational efficient and robust algorithms but requires an iterative process to remove the approximation errors. In this case, it turned out that two iterations are sufficient to converge at an accurate parameter set.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}' = \begin{bmatrix} 1 & -R_z & R_y \\ R_z & 1 & -R_x \\ -R_y & R_x & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}$$

3.3. Weight Compensated Motion Estimation (WCME)

Our approach to estimate the rigid-body motion is based on the continuous classification of model data into rigid and non-rigid movements. In order to achieve this goal, we have weighted the influence of the vertices used for the rigid-body motion estimation. The weights are associated to the Euclidean distance from the rigid body reference model to the current model. The idea is, that large deviations from the rigid-body constraint is caused by non-rigid deformation. We have used the $cos^2(x)$ function as weight function in the range between 0 and π. The Euclidean distance is scaled such that a weight of 0.5 is associated to the average distance of all vertices

classified as rigid-body.

$$\begin{aligned} w(i,n)\cdot\vec{v}_0(n) &= w(i,n)\cdot\left(\mathbf{R}\cdot\vec{v}_f(n)+\vec{t}\right) \\ f &\in \{1,...,F-1\} \end{aligned}$$

$$\begin{aligned} w(i,n) &= cos^2\left(\frac{e_{3D}(i,n)}{norm(i)}\cdot\pi\right) \\ norm(i) &= \frac{\bar{e}_{3D}(i)}{acos(\sqrt{0.5})}\cdot\pi \end{aligned}$$

Here, $w(i,n)$ represents the weight for each iteration i and for each vertex n. We have tested our Weight Compensated Motion Estimation (WCME) approach with a set of predefined manually labeled rigid-body vertices. A 3D face model is used, which is deformed along the y-axis from the eye-corners to the chin-tip, with a maximum of -50 mm at the chin-tip. The test model was rotated along the three axes in all combinations. The differences between the ground truth data and the estimated rotations are shown in **Tab. 1**. The remaining error is due to limited numerical accuracy.

method (two iterations)	**rotation angle** 1°	2°	3°	4°	5°
mean predefined	0.2	1.6	5.4	12.8	25.0
max predefined	0.3	2.2	7.4	17.5	34.2
mean WCME	28.0	28.0	28.9	32.1	37.2
max WCME	44.2	44.2	44.2	44.2	44.2

Table 1. Motion estimation test using predefined rigid-body vertices and our Weight Compensated Motion Estimation (WCME) approach. All errors in the table are specified in units of $[10^{-3}deg]$.

4. RECONSTRUCTION OF FACIAL DEFORMATIONS

Before the deformation can be analyzed, the rigid-body motion to the reference model has to be estimated and afterwards applied to the current model. Subtraction of the rigid-body motion results in the desired model deformation which also contains the measurement noise.

In order to effieciently respresent the vertex deformations caused by the facial motion, an Eigen vector decomposition of the 3D model sequence is used. For that purpose, all 3D coordinates are composed into a single matrix.

The three coordinates of all involved vertices (label visible) are placed alternatingly in the same column. The same technique is used for PCA of color images [15], where different color channel values are placed in the same column. This leads to a matrix $\mathbf{A}^{(3N,F)}$, where *N* belongs to the number of usable vertices and *F* to number of frames.

Because of the discrepancy between number of models and number of model vertices, the Eigen vectors are determined from the covariance matrix given by $\mathbf{A}\cdot\mathbf{A}^T$. Compared to the decomposition of $\mathbf{A}^T\cdot\mathbf{A}$, reduced computational power is required, because of the smaller size, which is three times the number of vertices instead of the number of frames. Both decompositions require also a different handling of the eigen vectors during reconstruction.

$$\mathbf{A}^{(3N,F)}_{reconstructed} = \mathbf{Eig}_c\cdot\left(\mathbf{Eig}_c^T\cdot\mathbf{A}\right)$$

In this equation, $\mathbf{Eig}_c$ are the Eigen vectors of the covariance matrix as defined above. Using only a subset of the Eigen vectors will reconstruct the data set with the smallest MSE compared to other linear transforms with same number of basis vectors. Thus the deformations in the face can be effiently described by a much smaller space. In our setup, the models consist of 127 vertices. Not all of these vertices are always visible, so that their positions cannot be specified correctly. Therefore, the maximum number of available eigen vectors in this scenario is 354.

5. FACIAL MOTION ANALYSIS

In this section, we present the results achieved by analyzing a sequence, where a person is counting from one to six.

Extracting the model deformation by subtracting the rigid-body motion with the first frame model as reference using our WCME approach provides us with a sequence of model deformations. The results of this extraction is shown in **Fig. 4**, where the upper line shows the Euclidean distance from the reference to the current model. This Euclidean distance has to be minimized by the rigid-body motion estimation. The vertical dashed lines represent the beginning of the spoken numbers. Between two spoken numbers the person tried to return to the first frame deformation, but an average deformation of around 0.5 mm remained. One result of the rigid-body motion estimation is that the almost similar performance of our proposed WCME approach compared to a predefined set of rigid-body motion vertices. Another interesting result is the distribution of the deformation

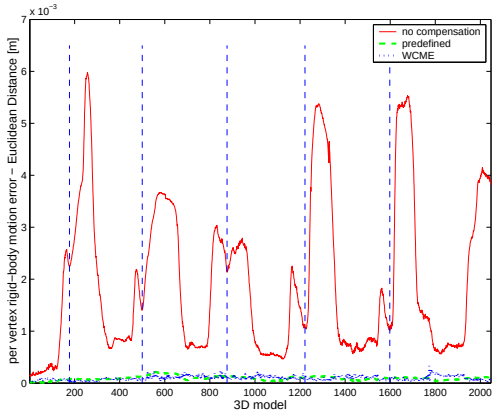

Fig. 3. Average vertex error caused by facial deformation. The different curves refer to different approaches for selecting vertices for the rigid-body motion estimation.

in a face. This result identifies the level of involvement of specific facial parts. Such a deformation map is displayed in **Fig. 4**, where the level of deformation is expressed as vertex diameter. Please note, that even the forehead vertices show some deformation. Therefore, these vertices cannot exclusively be used for the estimation of the rigid-body motion parameters. Analyzing the reconstruction quality of the used sequence leads to **Tab. 2** which gives us the desired association between the number of Eigen vectors and the resulting reconstruction error. Using 3 Eigen vectors leads to an average error of 0.5 mm whereas 8 Eigen vectors limit the maximum error to the same value. 23 Eigen vectors correspond to an average error of 0.5 pixels in the original image and 65 vectors are required to reduce the maximum error to 0.5 mm. Using this table and the visual results, given in **Fig. 5** for a frame model, allows the assumption of a high benefit for the compression of facial deformations in terms of Eigen vector linear combinations. The number of Eigen vectors required for a very good reconstruction can be limited to the first eight Ei-

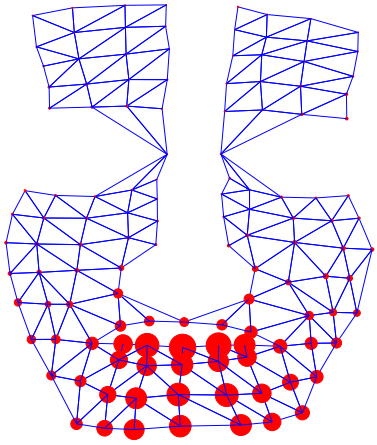

Fig. 4. Non-rigid body motion for each 3D model vertex. The circle located at each vertex describes the non-rigid body motion (high non-rigid vertex motion is equal to a big diameter of the circle).

	reconstruction error $[10^{-3}m]$					
max	3.03	1.56	1.02	**0.49**	0.23	**0.15**
mean	0.98	0.62	**0.49**	0.27	**0.15**	0.096
No.	1	2	3	8	23	65

Table 2. Reconstruction error: The bottom line depicts the number of Eigen vectors used for the reconstruction of the 3D model sequence. The columns above show the resulting Euclidean distance to the analyzed data set.

gen vectors, because the maximum reconstruction error using these Eigen vectors is smaller than 0.5 mm.

6. ACKNOWLEDGMENT

The work presented was developed within VISNET II, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 program. My special thank to Joachim Schüssler for his support in the mirror construction.

7. REFERENCES

[1] S. Perlman, "Contour Reality Capture System Unveiled," in *SIGGRAPH 2006*, July 2006.

[2] C. Trotman and J. J. Faraway, "Modeling facial movement: I. A dynamic analysis of differences based on skeletal characteristics," *Journal of Oral and Maxillofacial Surgery*, vol. 62, no. 11, pp. 1372–1379, November 2004.

[3] J. Faraway, "Modeling continuous shape change for facial animation," *Statistics and Computing*, vol. 14, no. 4, pp. 357–363, October 2004.

[4] I. A. Essa and A. P. Pentland, "Facial expression recognition using a dynamic model and motion energy," in *Proc. International Conference on Computer Vision (ICCV)*, Cambridge, MA, USA, June 1995, pp. 360–367.

[5] G. Kalberer and L. Van Gool, "Face animation based on observed 3D speech dynamics," in *Proceedings of the 14. Conference on Computer Animation*, Nice, France, November 2001, pp. 20–27.

[6] M. Odisio, G. Bailly, and F. Elisei, "Tracking talking faces with shape and appearance models," *Speech Communication*, vol. 44, pp. 63–82, October 2004.

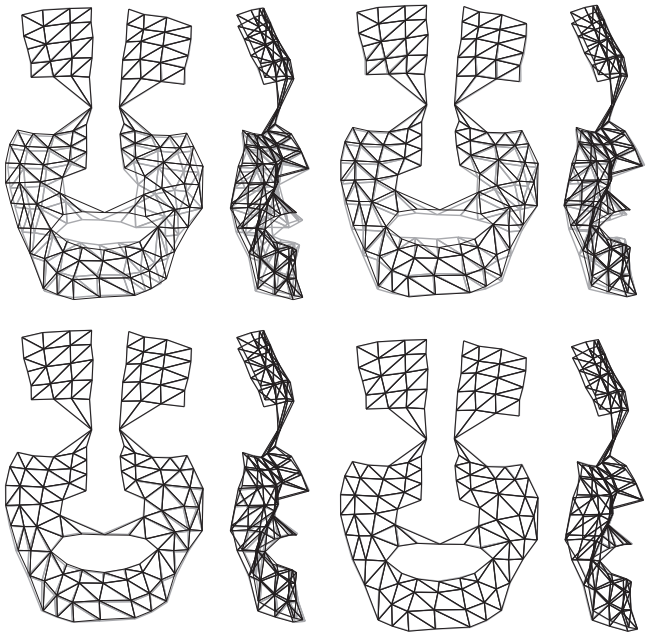

Fig. 5. Reconstruction results (gray belongs to the reconstructed model) at frame model 1975 using different numbers of Eigen vectors: (top left) one, (top right) two, (bottom left) three vectors with an average reconstruction error of 0.5 mm per vertex, and (bottom right) eight vectors with a maximum reconstruction error at 0.5 mm per vertex.

[7] S. Basu, "A Three-Dimensional Model of Human Lip Motions," M.S. thesis, Massachusetts Institute of Technology Department of EECS, Cambridge, MA, USA, February 1997.

[8] P. Eisert, "Model-based camera calibration using analysis by synthesis techniques," in *Proc. International Workshop on Vision, Modeling, and Visualization*, Erlangen, Germany, Nov. 2002, pp. 307–314.

[9] H. Li, P. Roivainen, and R. Forchheimer, "3-D Motion Estimation in Model-Based Facial Image Coding," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 6, pp. 545–555, 1993.

[10] P. Eisert and B. Girod, "Analyzing Facial Expressions for Virtual Conferencing," *IEEE Computer Graphics Applications: Special Issue: Computer Animation for Virtual Humans*, vol. 18, no. 5, pp. 70–78, September 1998.

[11] A. Yilmaz, K. Shafique, and M. Shah, "Estimation of Rigid and Non-Rigid Facial Motion Using Anatomical Face Model," in *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 1*, Washington, DC, USA, 2002, p. 10377.

[12] N. Ploskas, D. Simitopoulos, D. Tzovaras, G.A. Triantafyllidis, and M.G. Strintzis, "Rigid and non-rigid 3D motion estimation from multiview image sequence," *Signal Processing: Image Communication*, vol. 18, no. 3, pp. 185–202, March 2003.

[13] W. Huang, Y. Zhang, Y. Wang, and H. Cheng, "3D non-rigid motion estimation using the improved simulated annealing algorithm," in *4th International Conference on Machine Learning and Cybernetics (ICMLC 2005)*, August 2005, vol. 9, pp. 5330–5335.

[14] A. Del Bue, X. Lladó, and L. Agapito, "Non-rigid Face Modelling Using Shape Priors," in *Analysis and Modelling of Faces and Gestures (ICCV workshop), Lecture Notes in Computer Science. LNCS 2723*, Beijing, China, 2005, pp. 96–107.

[15] D. Alleysson and S. Süsstrunk, "Spatio-chromatic ICA of a Mosaiced Color Image," in *5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, 2004, vol. 3195, pp. 946–953.