ACTA STEREOL 1999; 18/2: 255-264
ORIGINAL SCIENTIFIC PAPER

EXTRACTION OF SEMANTIC 3D MODELS OF HUMAN FACES FROM STEREOSCOPIC IMAGE SEQUENCES

André Redert, Bart Kaptein, Marcel Reinders, Isabel van den Eelaart and Emile Hendriks
Information and Communication Theory Group
Faculty of Information Technology and Systems
Delft University of Technology
Mekelweg 4, 2625 CD Delft, The Netherlands
Tel +31 15 278 6269, Fax +31 15 2781843
Email: {andre,bart,marcel,isabel,emile}@it.et.tudelft.nl

Abstract

In this paper a system is described for the automatic generation of a specific semantic 3D model of a human face from a recorded stereoscopic image sequence. In the acquisition stage a calibrated stereoscopic camera records the specific face. After rectification of the image sequences a dense disparity field is estimated using a Markov Random Field approach. Next a triangle based 3D-wire frame model of the specific face is reconstructed by triangulation of corresponding left and right pixels. After compensation for rotation and translation, a triangle based 3D-wire frame model of a generic face is elastically matched to the specific face. As a consequence of this approach the resulting 3D model accurately represents the geometry of the subject with the availability of the face semantic information which is required for realistic representation and animation of human faces as well as for accurate facial expressions estimation.

Keywords: 3D-face modeling, stereoscopy, elastic matching, disparity estimation

1. Introduction

Within video coding, man-machine interfaces and 3D telepresence systems like virtual conferencing rooms, the modeling of human faces becomes increasingly important. Either because of its efficient representation (MPEG4 1998, SNHC 1997, Aizawa 1995), its life-likeness (Parke 1996, Thorisson 1997), or the required availability of 3D models (Hopf, 1994). In this highly active research field, most attention is paid towards either the tracking of facial expressions (Essa 1997, Yacoob 1996) or the recognition of faces (Turk 1991, Chellappa 1995). For realistic representations as well as for required expression estimation accuracy, it is essential to be able to automatically conform a known 3D generic face model (including a complete underlying muscle model) to the 3D structure of a specific person. The latter is, however, still an open problem. Often one falls back to manual approaches (Parke 1996, Waters 1991). Otherwise, one uses an automatic construction of 3D models from acquired stereo or range images (Braggins 1998), but leave out the mapping to the generic face model. This is, however, essential because the generic model provides information about the semantics (position of eyes and mouths etc.) and the underlying muscle structures (Waters 91). A few approaches exist that estimate the 3D position of a small set of landmark points (Bookstein, 1989), such as the corners of the eyes, and transform the generic model accordingly. A drawback of these methods is that they interpolate the 3D geometry of the specific person between the landmark points (Aizawa 1995, Nagashima 1991).

In this paper we present a new automatic conformation method that first accurately estimates the 3D geometry of a person and then conforms a generic 3D face model on the basis of the complete measured 3D data.

In section 2 an overview of the proposed system is given. Section 3 describes the acquisition of 3D data of a specific person and section 4 describes the conformation of the generic face model to the 3D data. Section 5 presents experimental results and in section 6 some conclusions are drawn.

2. System Overview

Figure 1 shows an overview of the proposed system. The system can roughly be divided into two subsystems, one that accurately estimates the 3D geometry of a recorded person and one in which a generic face model is conformed to this measured data.
To acquire 3D data we use the stereo imaging paradigm. A calibrated stereoscopic camera records the person. Here we have used a baseline between the cameras larger than the eye-distance to ensure high depth accuracy and a subpixel accurate calibration scheme. A 3D reconstruction of the scene is obtained by estimating the disparity between the left and right images.

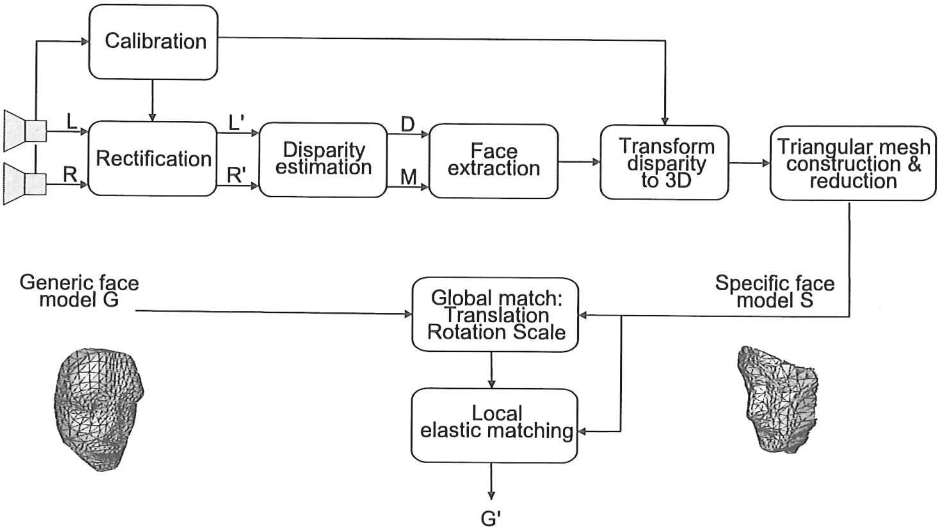

Figure 1: Overview of the proposed system for automatic generation of a specific and semantic 3D model of a human face

To capture the 3D data of the person's face first a 2D-face extraction scheme is applied. Here we exploit the fact that the face has a uniform color, has an elliptical shape (Parke, 1996) and that the face can be segmented in the disparity domain due to the restricted disparity range of the face. Next the disparity data of the extracted face is transformed into a collection of 3D points from which we construct a triangle based wire frame model.
In the second part a generic model is matched to this specific model. First, we perform a global matching based on the inertia properties of the models to correct for scaling, orientation and translation differences (Chaudhuri, 1991). Next, the generic model is deformed locally in an elastic way to match the specific model. The resulting model is a specific 3D model of the recorded human face in which semantic information is still available.
In the next sections, the different steps are described in more detail.

3. Specific Face Model Generation

This section describes the generation of the 3D model of the recorded face including data acquisition, camera calibration, disparity estimation, face extraction and surface reconstruction.

Data acquisition and camera calibration: By using the stereo paradigm, 3D data of the face of a subject is acquired. Two images are recorded with spatially separated cameras. By estimating the disparity, a 3D reconstruction of scene can be made. For an exact reconstruction of this scene the camera parameters such as orientation, position, and lens properties (focal length, lens distortion) must be known. These calibration parameters are also used to rectify the images in such a way that no vertical disparity occurs in order to ease the disparity estimation.

The cameras used are Panasonic E550 camera's with Fujinon TV Z lenses. The baseline of the cameras was about 20 cm. The calibration was done using a rectangular dark plate with 48 reflective circular markers. The coordinates of these markers are known with an accuracy of 0.02 mm. The camera calibration algorithm is described in (Sabel 1992) and (Woltring 1978). The rectification method was developed by R.J. Bakker (Bakker, 1997).

The procedure for camera calibration and image rectification is sketched in figure 2. Each time a specific face is recorded, the calibration plane is moved in front of the cameras and recorded in several different positions and orientations. The markers of the calibration plane are detected in the images, and their coordinates form the input for the calibration routine. The calibration routine calculates the camera positions, orientations, and lens properties iteratively using a least-means-square criterion. Using this information, two virtual camera's are constructed that are parallel and have no lens distortions, but with positions and lens properties as similar as possible to the real cameras.
To rectify the images virtual 3D mesh points are considered. Using the camera calibration information 2D reconstruction of these mesh points for each of the real and virtual cameras are determined. These resulting 2D meshes then form the input for a warping algorithm. This warping algorithm compares the 2D mesh of the real camera's with the 2D mesh of the virtual camera's and determines the new position of pixels of an input image in the rectified output image.

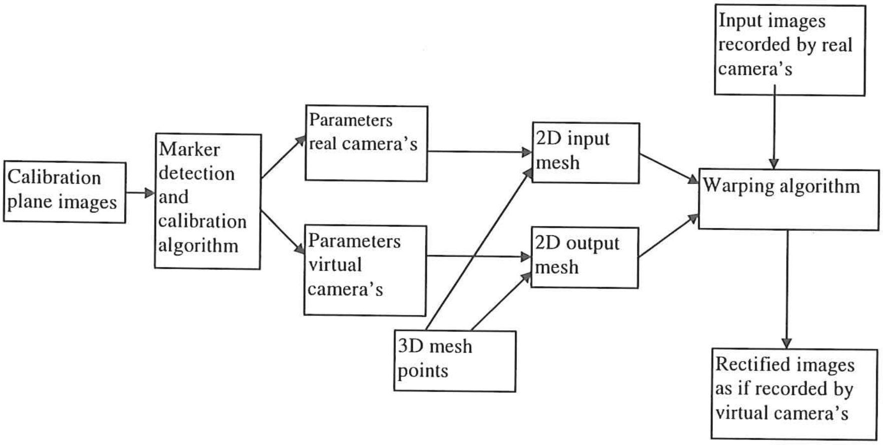
Input images recorded by real camera's
Calibration plane images
Marker detection and calibration algorithm
Parameters real camera's
Parameters virtual camera's
2D input mesh
2D output mesh
3D mesh points
Warping algorithm
Rectified images as if recorded by virtual camera's

Figure 2: Diagram of calibration and rectification routine.

Disparity estimation: Disparity estimation algorithms are based on luminance similarity of corresponding pixel pairs. These pixels pairs, however, will not always have exactly the same luminance value due to the noise, camera gain differences and specular reflectivity of the scene. Hence many factors influence the quality of the output field. Generally the resolution and accuracy is limited due to the algorithm (block, pixel of sub-pixel), the modeling of occlusion (Redert 1998, Stiller 1997), object orientation (Redert, 1998), object segmentation (Stiller, 1997) or the correctness of the luminance difference modeling.

To obtain a 3D specific face model with the best resolution and accuracy possible the disparity field should be dense (pixel resolution) and have (sub-) pixel accuracy. The modeling of occlusion and object segmentation is not taken into account since within the object of interest, the specific face, we do not want anything to be occluded. The non-occlusion requirement leads to the use of a small camera baseline. This has a positive side effect of minimizing luminance differences due to specular reflectivity on the subject's face which are difficult to model. However, this also implies that the range of disparity in the face is relatively low and sub pixel disparity accuracy will be necessary to increase 3D accuracy.

Popular approaches are block matching (Kanade, 1994) and pixel matching with dynamic programming (DP) (Redert, 1998) or with Markov Random Fields (MRF) (Stiller, 1997). Block matching is robust, but the relatively large size of the block does not allow for a pixel resolution disparity field. The DP algorithms allow for deterministic algorithms that yield the global optimum solution with easy incorporation of occlusion/segmentation. However, these algorithms are based on matching a single pair of scanlines in the images and thus they lack vertical consistency among scanlines. In MRF based algorithms the vertical consistency can be introduced leading to consistent high accuracy results (Stiller, 1997).

Here we adopt the MRF approach. The disparity values D(x,y) are defined as depicted in Figure 3. In stead of calculating the disparity values between left and right camera images, a virtual image exactly between both camera images is introduced, denoted as the center image.

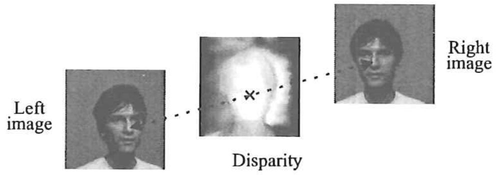

Figure 3: Center image containing disparity information.

Two energy terms are defined which sum has to be minimized. The first term is the external energy E_{ext} which minimization reflects the expected similarity in luminance difference $|I_L - I_R|$ for correct pixel pairs.

$$E_{ext} = \sum_{\substack{\text{all pixels}\\ \text{in center image}}} \left| I_L(x+D(x,y),y) - I_R(x-D(x,y),y) \right| \tag{1}$$

$$E_{\text{int}} = \sum_{\substack{\text{all neighboring}\\ \text{pixel pairs}\\ (x,y),(x^*,y^*)}} \mathrm{K} \left| D(x,y) - D(x^*,y^*) \right|^2 \tag{2}$$

The second term is the internal energy E_{int} which minimization smoothes the disparity field.

The disparity difference terms in (2) result from modeling disparity as a Gaussian Markov Random Field with 4-connected cliques. For more details about Markov Random Fields we refer to (Geman, 1984). The constant K influences the tradeoff between external energy and internal energy and is to be determined by experiment.

The disparity values D are real numbers to allow for sub pixel accuracy. The image coordinates in (1) are thus continuous. Image luminance values at non-grid positions are interpolated bi-linearly from the four neighboring grid positions.

For the minimization of $E = E_{int}+E_{ext}$ we use simulated annealing to avoid local minima. In simulated annealing (Geman, 1984), random perturbations are applied to all disparity vectors. If the energy E decreases, the perturbation is kept. If E increases, the perturbation is kept with a probability $e^{-\Delta E/T}$, governed by a temperature T. The temperature is decreased slowly to lead to convergence of the algorithm. To speed up convergence a hierachical approach is adopted. In the hierarchical approach a pyramid of images is constructed from both left and right images. The algorithm starts at the lowest detail level (images are 2*2 pixels) with the disparity field initialized to zero. At this level, N_{start} iterations are performed in which each disparity value is visited once. At each visit, a perturbation is selected from a uniform distribution between -2 and 2, the temperature is equal to T_{start}. At each transition to a higher detail level, the disparity field from the lower level is interpolated linearly to the higher level and used as starting point. The temperature is divided by 2, the number of iterations is divided by 4. The start temperature and number of iterations are to be determined experimentally.
From the estimated disparity, an interpolated center image is constructed, based on the left and right images (Redert, 1997) which then is used for face extraction.

Face extraction: Since we are only interested in the pixels corresponding to the face we have developed a 2D-face extraction scheme. Here we exploit the fact that the face has a uniform color and has an elliptical shape. This results in an elliptical shaped mask, which we used to segment the disparity image. In the disparity estimation stage errors were introduced because of the smoothing of the data. In fact the face is glued to the background, which cause wrong disparity data especially near the contour of the face. To correct for these errors we threshold the segmented disparity data to obtain that part of the face between the point of the nose and roughly the beginning of the ears.

Surface reconstruction: Because the 3D coordinates are organized in the image grid, surface reconstruction can be done easily by defining two triangles between each set of four adjacent node points. A triangular surface reconstruction is necessary because the matching method is based on triangular surface models. The resulting triangular surface model consists of node points, or vertexes, connected by edges to form triangular surface patches or polygons.
To speed up the matching procedure, the number of triangular surface patches is reduced using Jade, a multi resolution decimation based on a global error (Ciampalini, 1997).

4. 3D Model Matching

This section describes how a generic 3D model of the face is deformed elastically to match the obtained 3D model of the specific face. The elastic surface matching is an extension of the elastic matching of line drawings, described by Burr (Burr, 1981) to three dimensions.
The matching procedure consists of a global matching step (alignment and scaling) and a local matching step.

Alignment and scaling using moments of inertia: To find maximum correspondence between the generic and specific model, the models are first aligned and scaled as well as possible. For this, different transformation methods can be used. We used a method based on the inertia properties of the models (Chaudhuri, 1991). The principal axes of the models are calculated, and the bounding boxes, aligned with these principal axes, are defined. Using these bounding boxes, a translation vector $\boldsymbol{t}$, rotation matrix $\boldsymbol{R}$, and scaling vector s, are calculated to transform the generic model such that it matches the specific model as well as possible.

Local matching by deformation of the generic model to match the specific model: After alignment and scaling, geometrical differences between the generic and specific model are diminished by an elastic transformation of the generic model. This transformation tries to minimize the dissimilarity measure between both models which is defined as the mean distance of all the node points of the generic model to the surface of the specific model and vice versa.

The mean distance is calculated with a non-uniform weighing factor, so dissimilarity in the direction of the frontal plane is weighed more than distances perpendicular to this plane.
Using this dissimilarity measure a 3D force field can be defined acting between the generic model M_G and the specific model M_S. This force field consists of two parts: the pulling vectors f_{pull} defined as the vectors pointing from their corresponding projection points $p_{pS}(j)$ on M_G towards the points $p_S(j)$ of M_S. The pushing vectors f_{push} are defined as the vectors from the points $p_G(i)$ of M_G pointing towards their corresponding projection points $p_{pG}(i)$ on M_S (see figure 4).

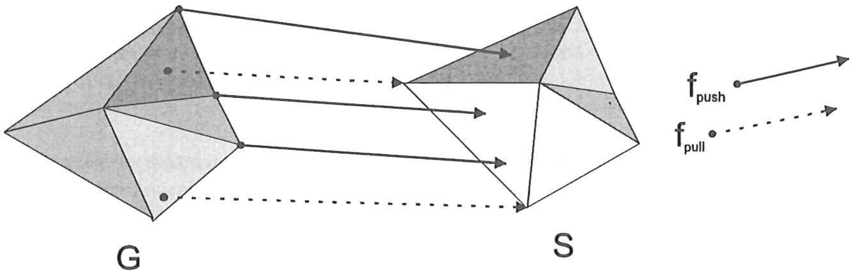

Figure 4: Part of the force field between M$_G$ and M$_S$.

Based on these sparsely defined pulling and pushing vectors, a smooth displacement field $D(p)$ can be defined for any arbitrary point p in the 3D space. This is achieved by calculating a weighed average of the sparse force vectors $f_{pull}(j)$, and $f_{push}(i)$, such that force vectors close to point p have more influence than force vectors further from p. This weighing behavior is modeled by taking the Gaussian of the distance between p and the base of the force vectors:

$$D(p) = \frac{1}{2\gamma}\left(\frac{\sum_{i=1}^{N_G} G_G f_{push}(i)}{\sum_{i=1}^{N_G} G_G} + \frac{\sum_{i=1}^{N_S} G_S f_{pull}(i)}{\sum_{i=1}^{N_S} G_S} \right) \tag{3}$$

where G_G and G_S are defined by:

$$G_G = \frac{\exp^{\frac{-\|p-p_G\|^2}{2\sigma_s}}}{\sigma_s\sqrt{2\pi}} \tag{4}$$

$$G_S = \frac{\exp^{\frac{-\|p-p_{pS}\|^2}{2\sigma_s}}}{\sigma_s\sqrt{2\pi}} \tag{5}$$

N$_G$ and N$_S$ are the numbers of vertices of the generic model and specific model, respectively.
The factor γ controls the damping. Values less than 1 result in overshoot, and values greater than 1 result in undershoot.
Here σ_S controls the size of the neighborhood in which force vectors still affect the displacement of p. In effect σ_S is a very important parameter of the proposed method because it controls the elasticity of the generic model. When σ_S is very large, all points will be displaced, while small σ_S will cause very

localized displacements. σ_S is the stiffness parameter at the s^{th} stiffness stage. Burr proposed to change σ_s iteratively according to:

$$\sigma_s = \sigma_0 F^{-s} \qquad s = 0,1,2,3,.... \tag{6}$$

where F is a constant $1 < F < 2$, (we used 1.2). The value σ_0 depends on the size and shape of the models to transform.

To reach maximum correspondence between the generic and specific model, the geometric differences should be removed very slowly. This implies that deformations, as caused by the smoothed displacement field in eq. (3), to the source model are applied in an iterative way. One such deformation is defined by:

$$W^k = W^{k-1} + D^{k-1} \qquad k = 0,1,2,3,.... \tag{7}$$

$$W^0 = M_G \tag{8}$$

Where W^k represents the "warped" generic model after k iterations of elastic matching. During the iteration process, the stiffness of the model σ_S decreases, so deformations of the model become more local. To increase the robustness of the method, we made the stiffness at iteration k, depending on the maximum displacement d_{max} of the model at iteration k-1.

$$d_{\max} = \max_{\forall(\bar{p} \in M_G)} \{\|D(\bar{p})\|\} \tag{9}$$

As a result of this, the model goes to the next stiffness stage only when the displacements in the current stiffness stage become smaller than a factor ξ of the current stiffness. This extra speed-limiting factor is built in to make sure that the stiffness of the model does not decrease too fast.

$$\text{IF } (d_{\max}^{k-1} < \xi\sigma_s^{k-1}) \text{ THEN } s = s+1 \tag{10}$$

With factor ξ between 0 and 1 (we used 0.2). The stopping criterion is defined as:

$$d(M_G, M_S) < threshold \tag{11}$$

The parameter *threshold* can be chosen as small as required (at the cost of processing time).

5. Experimental Results

We recorded several test sequences under different lighting conditions. Figure 5 shows the recorded stereo image pair in weak diffuse lighting. Figure 6 illustrates one of the images used for calibration. The rectified images are shown in figure 7.

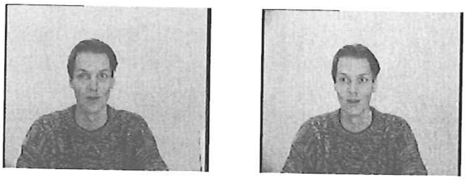

Figure 5: Original image pair

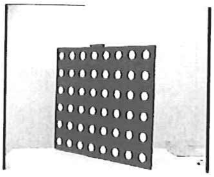

Figure 6: Image of calibration pattern

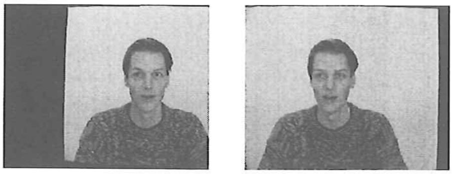

Figure 7: Rectified image pair

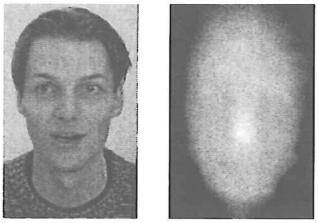

Figure 8: Center image and disparity field

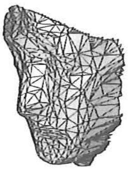

Figure 9: Specific Face

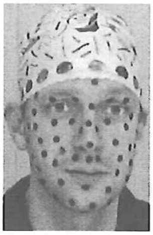

Figure 10: Recorded image with texture added to the subject's face

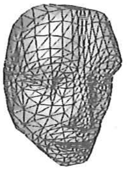

Figure 11: Generic Face

In figure 8 a (part of) the center image and resulting disparity field can be seen. After face segmentation and surface mesh generation the specific face S is obtained, shown in figure 9. Figure 10 illustrates an image with texture added to the face that was used to test the matching stage of the algorithm. Figure 11 shows the generic face G. Finally, figure 12 shows three models in three viewing positions. The left-most model is the reconstructed specific face of figure 10, the right-most model is the globally matched generic face according this data and the center model is the locally matched face and final result.

The reconstructed specific face clearly resembles a face. When comparing the eye distance between the real subject and the reconstructed specific face they both approximately measure 6.5 cm. The eyes and nose are clearly visible, however, the mouth has very little depth details. Throughout all of the face, a small amount of noise is present due to disparity estimation errors that were in the order 1.0 pixel. These errors resulted from camera luminance gain differences and specular reflectivity of the subject's face and, if necessary, can be reduced by using active structured light or additional texture (see figure 10).

The shape of the chin and length of the nose of the resulting face (locally matched) clearly matches

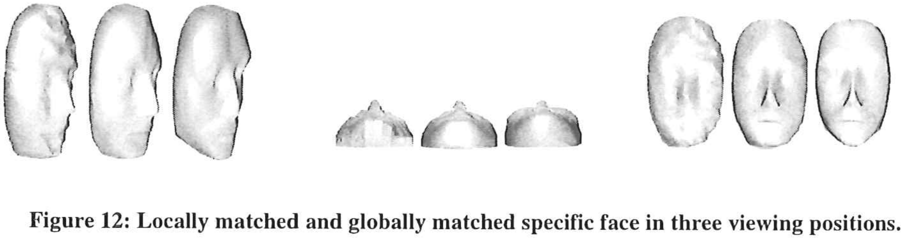

Figure 12: Locally matched and globally matched specific face in three viewing positions.

the specific face. Additionally, it has inherited the smooth face features from the generic model.

We compared the three models using a dissimilarity measure defined as the average distance between the three models in figure 12. The specific and generic face have average distance 10.2 mm, the resulting and generic face have distance 9.8 mm, the specific and resulting face have distance 1.8 mm. The subjective and objective evaluations show that the matched face contains the smoothness and clear features of the generic face together with the geometry of the specific face.

6. Conclusions

In this paper we have presented a system for automatic generation of a specific and semantic 3D model of a human face from a recorded stereoscopic image sequence. We have shown that the system is capable of producing good results. The resulting 3D face model accurately represents the geometry of the recorded subject whilst still containing the semantic information and underlying muscle structure. The results can be further improved by optimizing the recording. That is, either improving the lighting conditions during image acquisition or use of structured light, multiple cameras or other modalities (e.g. range data). The elastic surface matching stage is very promising. Although a good initial global matching is required (provided by the matching of the moments of inertia) the elastic matching is able to deform the generic model in such a way that the semantic information (mouth position, eye position, etc.) is transferred to the specific model.

REFERENCES

Aizawa, K., Choi, C., Harashima, H. and Huang, T.S. *Human facial motion analysis and synthesis with application to model-based coding.* In Sezan, M. and Lagendijk, R., editors, Motion Analysis and Image Sequence Processing, Kluwer Academic Press, 1993: chapter 11, pages 317-348.

Bakker R.J., *"Rectification of Stereo Images on Behalf of Disparity Estimation"*, M.Sc. Thesis, Information Theory Group, Delft University of Technology, 1997.

Bookstein F.L., *Principal Warps: Thin Plate Splines and the Decomposition of Deformations*, IEEE Trans. On PAMI, 1989: 11(6).

Braggins, D. (1998). *What to watch for in Euro Machine Vision*, 1998. Advanced Imaging, feb, 1998: page 16. (http://www.turing.gla.ac.uk)

Burr D. J., *"Elastic Matching of Line Drawings*", IEEE Trans. on PAMI, 1981: 3(6); 708-713.

Chaudhuri B.B. and Samanta, G.P. *"Elliptic fit of objects in two and three dimensions by moment of inertia optimization"*, Pattern Recognition Letters. 1991: 12; 1-7.

Chellappa, R., Wilson, C.L., and Sirohey, S. *Human and Machine Recognition of Faces: A Survey.* Proceedings of the IEEE, 1995: 83(5); 704-740.

Ciampalini A. Cignoni P., Montani C., and Scopigno R.. *Multiresolution Decimation based on Global Error*, The Visual Computer, Springer Verlag, 1997: 13(5),
(Jade: http://miles.cnuce.cnr.it/cg/homepage.html)

Essa, I.A., and Pentland, A.P. *Coding, analysis, interpretation and recognition of facial expressions.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997: 19(7);757-763

Geman S. and Geman D., "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Transactions on PAMI*, 1984: 6(6); 721-741.

Hopf, K., Runde R. and Bocker M., *"Advanced videocommunications with stereoscopy and individual perspective"*, In Towards a Pan-European Telecommunication Service Infrastructure, IS&N '94, Kugler et al., Berlin, Heidelberg, New York, Springer, 1994.

Kanade T. and Okutomi M., "A stereo matching algorithm with an adaptive window: theory and experiment", *IEEE Transactions on PAMI*, 1994: 16(12); 1207-1212.

MPEG-4: ISO/IEC-JTC1/SC29/WG11 N2196 (1998). *MPEG-4 Overview, Coding of moving pictures and audio.* International Organization for Standardization. 1998 (http://drogo.cselt.stet.it/mpeg/public/w2196.htm)

Nagashima, Y., Agawa, H. and Kishino, F. *3D Face model reproduction method using multi view images.* In proceedings SPIE Visual communication and image processing, 1991: 1606;566-573.

Parke, F.I., Waters, K. *Computer Facial Animation.* A.K. Peters Ltd., Wellesley, MA., 1996. ISBN 1-56881-014-8.

Redert P.A., Hendriks E.A., Biemond J., *"Synthesis of multi viewpoint images at non-intermediate positions",* Proceeding IEEE Int. Conference on Acoustic, Speech and Signal Processing, 1997: IV; 2749-2752.

Redert P.A., Tsai C.J., Hendriks E.A. and Katsaggelos A.K., *"Disparity estimation with modeling of occlusion and object orientation", Proceedings of VCIP98,* pp. 798-808, 1998

Sabel J.C., *"Implementation of SMAC in a 3D motion Analysis system",* Motion Studies Lab, Delft University of Technology, internal document, 1992.

Stiller, "*Object-based estimation of dense motion fields*", IEEE Transactions on Image Processing, Vol. 6, No. 2, pp. 234-250, 1997.

SNHC-97 ISO/IEC-JTC1/SC29/WG11 N1669m3. *MPEG-4 SNHC, Coding of moving pictures and audio.* International Organization for Standardization, 1997. (http://drogo.cselt.stet.it/mpeg/faq/faq-snhc.htm)

Thorisson, K.R. Communicative Humanoids*; A Computational Model of Psychosocial Dialogue Skills.* Ph.D. Thesis. School of Architecture & Planning, Massachusetts Institute of Technology, 1996

Turk, M. and Pentland, A. *Eigenfaces for recognition.* Journal of Cognitive Neuro Science, 1991: 3(1);71-86.

Waters, K. and Terzopoulos, D. *Modeling and animating faces using scanned data. The journal of visualization and computer animation*, 1991: 2;123-128.

Woltring H.J., *"Simultaneous Multi-frame Analytical Calibration (SMAC) by recourse to oblique observations of planar control distributions",* SPIE Applications of Human Biostereometrics, 1978: 166; 24-135.

Yacoob, Y. and Davis, L.S. *Recognizing human facial expressions from long image sequences using optical flow.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996: 18(6);636-642.