3D-Reconstruction of Faces: Combining Stereo with Class-based Knowledge

Christian Wallraven, Volker Blanz, and Thomas Vetter

Max-Planck-Institute for Biological Cybernetics,
Spemannstrasse 38, 72076 Tübingen, Germany,
christian.wallraven@tuebingen.mpg.de,
http://www.kyb.tuebingen.mpg.de/

Abstract. The recovery of the threedimensional structure of faces with conventional stereo methods still proves difficult. In this paper we introduce a higher order constraint based on linear object classes, which supplies a standard stereo algorithm with prior knowledge of the general structure of faces. This constraint has been learned by exploiting the similarities between 200 faces in a database and is represented in a morphable face model.

This combined approach has been tested and compared against an already existing method for estimating depth information using only prior knowledge and against the standard stereo algorithm.

1 Introduction

The reconstruction of a threedimensional model of a face from two (or more) images has interesting applications in many areas such as face recognition, video conferencing or even in the film industry. This threedimensional model can be obtained with a stereo approach. Stereo is a vision problem that has recieved much attention over the past few decades. Since the Marr and Poggio classic paper [10] numerous other approaches for solving this difficult problem have been proposed. To follow ([5, 12], see also [3] for a good overview) one can divide these into two broad main categories: intensity-based and feature-based methods.

Intensity-based methods measure pixel properties and correlate these to find corresponding points. There are approaches matching single pixels [2] and others matching windows consisting of a small patch around the pixel [4, 9]. Two pixels are said to be matched when a maximum in correlation is found.

Feature-based methods first process the images to extract higher order features such as lines, edges, or even surfaces which are then matched. This is done in order to provide a very reliable and robust estimation of depth, which is less sensitive to noise. Both types of approaches need to impose constraints to solve this highly ambigous problem. Typical constraints are:

1. epipolar constraint: based on the projective geometry corresponding points must lie on epipolar lines
2. uniqueness constraint: every point has at most one corresponding match

3. ordering constraint: the order of the matches is the same in both images
4. smoothness constraint: the change in disparity[1] is limited

The last three constraints are lower order constraints which are valid for a wide range of objects (opaque and 'smooth' objects). One of the situations which proves difficult for stereo is - as mentioned - the reconstruction of faces. First, these differ widely between individuals in properties such as texture and shape and second, for a given face, properties such as the reflective behaviour of the skin change over the whole face. 'Classic' features as lines are only present in very prominent facial areas such as the mouth, the eyes and wrinkles or resulting from illumination at the edges of shadows. But of course this information can only be used for very sparse depth reconstruction. Due to the lack of texture on the other hand, intensity-based algorithms which are capable of producing dense disparity data usually fail in areas such as the cheeks or the forehead. In addition, scenes with a large range of disparity make the search for correspondence more difficult as more 'similar' points can be found within the search range.

In order to overcome these difficulties it seems natural to impose higher order constraints to facilitate the search. This can for example be prior knowledge about the specific shape of the objects which are to be reconstructed. To go one step further one can base the search for corresponding points on a generic model of the objects. In recent years a method was developed which is purely based on prior knowledge and is able to estimate the 3D-structure of a face given only a single image [1, 13]. This method uses a morphable face model which was learned from a database of 200 faces. This morphable model is first fitted to the two stereo images and then serves as a guideline for the search for correspondences in a conventional stereo algorithm. There are similar approaches with a face model (most notably [7] and [8]); however, these models are restricted to points on a wire-mesh, whereas our model is capable of producing very dense and flexible shape and texture data.

For the rest of the paper we will proceed as follows: First, we present the outline of the algorithm we use. The two main parts of the algorithm - the fitting of the model to a stereo image pair and the stereo algorithm running on this data - are described. We will touch briefly on the problems encountered when putting the result of the stereo algorithm back into the model. In the third part the results we obtained with this approach are presented, and the final part summarizes these and gives an outlook on future work.

2 Stereo with Prior Knowledge

First an overview of the algorithm we use (see Fig.1):

We first fit the linear face model to the two stereo images. Then a standard stereo algorithm takes the resulting disparity map as input and corrects it. The output disparity map is then triangulated to obtain the 3D-data of the face.

[1] In the simplest case disparity is just the x-coordinate difference for corresponding points in the two images.

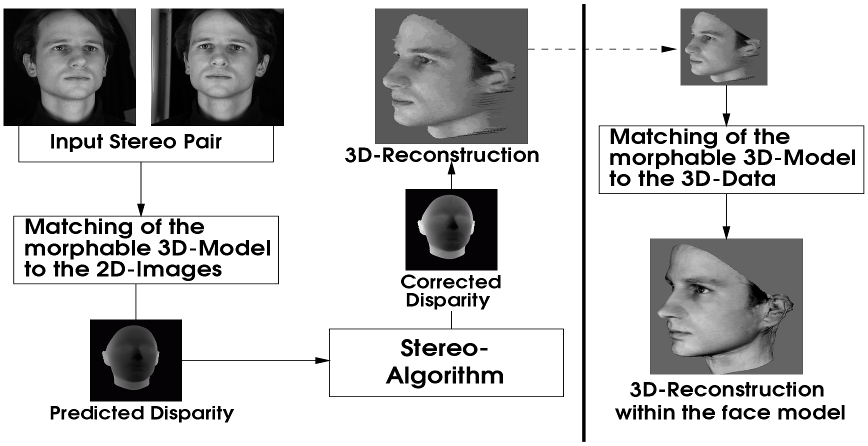
Input Stereo Pair
Matching of the morphable 3D-Model to the 2D-Images
Predicted Disparity
Stereo-Algorithm
Corrected Disparity
3D-Reconstruction
Matching of the morphable 3D-Model to the 3D-Data
3D-Reconstruction within the face model

Fig. 1. Overview of the Algorithm. The left part shows the combination of a standard stereo algorithm with our morphable face model. The right part shows the fitting of the reconstruction back into the model.

In addition it is possible to fit the morphable model to this threedimensional reconstruction to obtain a model-based representation of the face (Fig.1, right).

The test-data consisted of pictures of 14 persons (8 men and 6 women) which we took with two off-the-shelf high resolution digital cameras. The resulting images are of size 1024x1024 pixel (due to the camera architecture this is an interpolated resolution) and thus able to show very fine details all over the face. To test our approach, the persons were scanned immediately after taking the two stereo pictures by a Laser - Scanner (CYBERWARE^{TM}), which provides a very accurate threedimensional model of the face. This scanner was also used to generate the face database from which the linear model was learned [1, 13].

2.1 Fitting a Linear Face Model to a Pair of 2D Images

The morphable face model exploits prior knowledge from a database of 200 laser scans of human faces (not containing the 14 test faces), and explicitly captures the range of variations that occurs within this class of objects [13]. Each face is represented by a shape vector $S = (X_1, Y_1, Z_1, ..., Z_n) \in \mathbb{R}^{3n}$, that contains the X, Y, Z-coordinates of its n vertices, and a texture vector $T = (R_1, G_1, B_1, ..., B_n) \in \mathbb{R}^{3n}$, that contains the R, G, B color values of the same vertices. The morphable model is defined as the shape space spanned by $\sum_{j=1}^{200} a_j S_j$, and the texture space spanned by $\sum_{j=1}^{200} b_j T_j$ of all face prototypes j. Linear combinations within shape and texture space will describe new possible faces only if correspondence between all prototypes has been established. Corresponding points on each of the prototype faces, such as the tip of the nose, have to be described by the same vector component i in all vectors S_j and T_j, which is achieved by a gradient-based optic flow algorithm [13].

For any set of model parameters a_j and b_j, and head orientation, position and illumination parameters, we can compute a 2D image I_{model}. In a gradient descent algorithm, these parameters can be optimized such that the difference between I_{model} and a given input image I_{input} is minimal [1]. In a similar way, the model can be matched to several images $I_{k,input}$ simultaneously by using independent variables for position and orientation in each view k, and minimizing the sum of image differences $||I_{k,model} - I_{k,input}||$.

2.2 The Stereo Algorithm

A simple intensity-based stereo algorithm is used. To find corresponding points in the left and right image a window is placed at a position x and an according window is then slid along the epipolar line in the right image. For each disparity value d the normalized cross-correlation between these two windows is evaluated:

$$NCC = \frac{\sum_{k=x-w}^{x+w} \sum_{l=y-w}^{y+w} (I_1(k,l) - \overline{I}_1) \cdot (I_2(k+d,l) - \overline{I}_2)}{\sigma_1 \cdot \sigma_2} \quad (1)$$

where $2(w+1)$ is the width and height of the window, $\overline{I}_{(1,2)}$ and $\sigma_{(1,2)}$ are the mean and the variance of the intensity data in the left and right window, respectively. The use of the normalized cross-correlation has the advantage that intensity differences in the pictures due to different illumination can be accounted for. A second advantage is that the calculation of the variance for the window provides an estimate of how much change in the intensity values there is, which relates to the amount of 'texture' in the window.

For a robust estimation of disparity the window size has to be large enough in order not to be disturbed by the image noise. On the other hand, it has to be small enough to accurately capture the finer details in the images and to take into account the assumption of a frontoparallel surface with constant disparity in this window. The second point stems from the fact that for simplification purposes the search windows are assumed to be rectangles, whereas the correct form of the window in the right image is an affine transformation of a rectangle depending on the surface normal at the given point (e.g. [6]).

To increase robustness and reduce calculation time a hierarchical matching strategy is employed. Since the disparity range we are working with is rather large (well over 130), some outliers still exist. To reduce these a second run of the stereo algorithm is made, but this time with the left and right image swapped. Since the cross-correlation formula (1) is asymmetric with respect to the intensity values I_1 and I_2 only disparity values which are the same in both directions are taken to be correct (this can be seen as an explicit implementation of the uniqueness constraint). This stereo algorithm also needs initialization of the search range (which is available via the camera geometry) to avoid massive outliers and to further reduce calculation time.

2.3 Combining Stereo with the morphable Face Model

In the first stage, the stereo algorithm uses the predicted disparity map of the model to reduce its search range. The advantages are that:

1. the calculation time decreases
2. no prior initialization is needed since this is supplied by the model
3. a disparity value can be assigned to a higher number of points
4. the number of outliers is reduced, thus increasing accuracy
5. a dense disparity map is obtained, since in regions where stereo fails the model estimates can be used

The resulting disparity map is then triangulated to obtain the 3D-reconstruction.

Since in some cases the changes made by the stereo algorithm are quite drastic with respect to the first model estimate, the resulting 3D-reconstruction has bumps and peaks in it. To obtain a smooth face which again is part of the span of the model, the correspondence between the 3D-data and the morphable model has to be computed. For this we tested a similar strategy as outlined in section 2.1, whereas this time the matching process is calculated on a parametrized surface [1]. Currently, this algorithm matching the morphable 3D-model to the 3D-data smoothes the data too much, but we hope to improve on that in the future. Some of these additional improvements will be outlined in the last section.

3 Results

The ground truth for the images is only approximately available (the Laser-scans were taken right after the two stereo images) due to slight changes in expression and pose of our subjects. For this 'ground truth' we took the disparity image resulting from a model matching process of the *original* Laser-scan to the corresponding stereo images. We then calculated the mean disparity error per pixel from the difference of this disparity map and the disparity maps obtained

1. using only prior knowledge, i.e. the morphable model
2. using only the stereo method
3. after running the stereo algorithm on the model data

where only points with disparity values in all three disparity maps were taken into consideration. It has to be said that this error measure should not be taken as the 'absolute' truth, since the matching of the scan to the two images itself is prone to errors. Furthermore, slight rotations of the face result in drastic changes in disparity in regions with high disparity gradient (such as the chin or the jaw line), so that the errors tend to be biased by this effect. Rather, these values should give a general idea of the performance of the three algorithms. We are using the disparity error measure because an error estimate in 3D-space would - again - require the calculation of the full correspondence between the ground truth model and the stereo data.

The table below lists the mean disparity error per pixel for all subjects showing that stereo alone improves on the model prediction while the combined approach yields the best overall performance. These results are discussed in the next three sections.

	Model	Stereo	Stereo+Model
disparity error per pixel	7.72	6.06	5.35

3.1 Only Prior Knowledge

The disparity map as predicted by fitting the linear face model to the two stereo images (independently) is shown in Fig.2a. This is of course very smooth and provides a disparity value at each point in the face. All model predictions have in common that - due to a lack of a suitable constraint when fitting to *two* images - the resulting profile of the 3D-Model is only an estimate (see also Fig.3a), so that this method yields the highest error.

3.2 Only Stereo

As expected the algorithm performs better on the relatively well textured faces of the male subjects (mean of accepted points for male subjects: 35.6 percent, for female subjects: 32.5 percent). In Fig.2b[2] the reconstruction of the stereo pair of Fig.1 with the stereo algorithm without consistency check is shown. The reconstruction suffers -as expected- from the typical artifacts. First, the corona effect occurs at depth boundaries. This is an effect due to the window-matching method, where the position of the disparity maximum at object boundaries in the worst case is moved half the window size [4]. The multi-resolution approach causes a further smearing of the boundary from one level to the next (see region around the chin in Fig.2b). Second, if the search range for the coarse-fine algorithm is too large, it tends to run into other local minima leading to outliers, as similar points exist in many regions of the face. Third, points in the left image which are hidden in the right image are of course assigned arbitrary disparity values (see for example black patch at the left ear in Fig.2b). The consistency check removes most -but not all- of these errors (Fig.2c). To obtain dense data for reconstruction this disparity map has to be smoothed and interpolated. This was done by putting the estimated points into a physical spring- or membrane-model (corresponding to a quadratic potential function), where a global minimum in

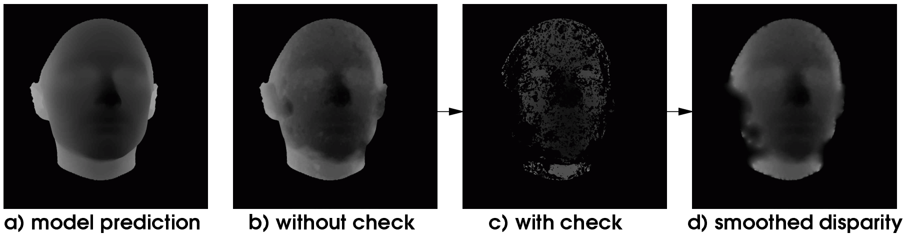

Fig. 2. The model provides a dense estimate. Without the consistency check stereo produces many outliers and false matches; the check mostly eliminates these. The smoothed disparity map supplies disparity only in some areas.

[2] to facilitate comparison the stereo data is clipped by the model data from Fig.2a; the disparity maps encode depth with intensity - the lighter the farther away

the potential energy of the model is sought. The result is shown in Fig.2d. It is obvious that in regions with high change in disparity (ears) or regions with insufficient texture (neck, nose) no reliable estimation can be obtained so that these remain undetermined. The error of this method is lower than with the model-only approach, but due to still existent outliers not optimal.

3.3 Combined Approach

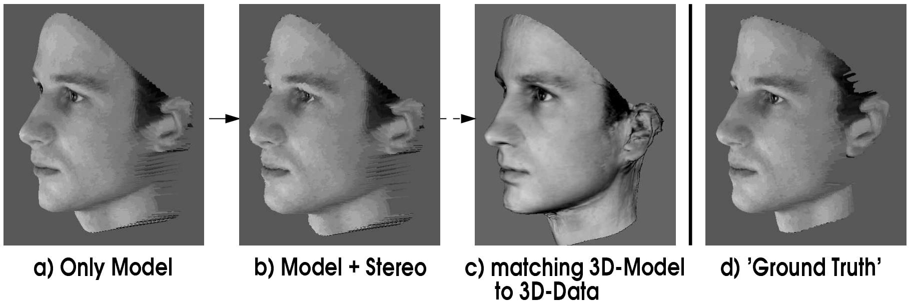

Fig. 3. Comparison of the model prediction, the combined approach and the algorithm matching the model to the 3D-data with the original scan.

Again the algorithm performs better on the male subjects (mean of accepted points for male subjects: 38.3 percent, for female subjects: 36.0 percent, which is ≈3 percent more than with stereo alone). In Fig.3a we show the reconstructed face as predicted by the model approach, Fig.3b is the reconstruction after performing stereo on this data. There is considerable improvement - especially in the profile - if you compare both faces to the 'ground truth' data (Fig.3d). This method improved all of the model predictions considerably, which we checked visually by comparison with the 14 original scans (such as in Fig.3d) and which is reflected in the lowest error of all three approaches.

Fig.3c shows the face after the matching algorithm fitting the 3D-model to the 3D-data at a higher resolution. Notice that in this image the *full* shape was predicted by the model. Here you can see that again a smooth face with less peaks is obtained, but at the price of countermanding some of the changes made by the stereo algorithm (e.g. the mouth region and the nose).

4 Summary and Outlook

We presented a novel stereo algorithm for the reconstruction of faces that incorporates a higher order constraint based on a linear object class model. By combining stereo with the model prior knowledge about the face is used, which makes the search for corresponding points in the two stereo images more robust and results in a dense estimation of disparity. By comparing the results obtained with this new approach with the stereo-only or model-only approach we could show significant improvements. The comparison with stereo shows that the accuracy of the disparity estimation is improved by reducing the numbers of outliers and that a higher number of points can be put into correspondence. The comparison with the model-only approach shows that the combined approach successfully corrects the error in the threedimensional estimation of the face.

Since the combination of the resulting stereo data with the model proved difficult, we hope to improve on that in future work. To further improve the results of the algorithm matching the 3D-model to the 3D-data an iterative technique would be possible, where at each iteration step the stereo data is 'slowly injected' into the model and then the matching is done. Another possibility we are pursueing is an approach following [11], where a dense reconstruction can be achieved via a local interpolation of the sparse disparity data (such as in Fig.2b) generated by the stereo algorithm.

References

1. V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. *Computer Graphics Proceedings SIGGRAPH'99*, 1999. Accepted.
2. I. J. Cox, S. L. Hingorani, B. M. Maggs, and S. B. Rao. A Maximum Likelihood Stereo Algorithm. *Computer Vision and Image Understanding*, 63(3):542-567, 1996.
3. U. R. Dhond and J. K. Aggarwal. Structure from Stereo - A Review. In *IEEE Transaction on Systems, Man, and Cybernetics*, 19(6):1489-1510, 1989.
4. L. Falkenhagen. Hierarchical Block-Based Disparity Estimation Considering Neighbourhood Constraints, *International Workshop on SNHC and 3D Imaging*, 1997.
5. O. Faugeras. Three-Dimensional Computer Vision. A Geometric Viewpoint. The MIT Press, Cambridge, 1993.
6. O. Faugeras and R. Keriven. Complete Dense Stereovision using Level Set Methods. In H. Burkhardt, B. Neumann, eds., *Computer Vision - ECCV'98, Vol.I*, 379-393, 1998.
7. P. Fua and C. Miccio. From Regular Images to Animated Heads: A Least Squares Approach. In H. Burkhardt, B. Neumann, eds., *Computer Vision - ECCV'98, Vol.I*, 188-202, 1998.
8. S. B. Kang. A Structure from Motion Approach using Constrained Deformable Models and Appearance Prediction. *Digital Equipment Corporation, Cambridge Research Lab, Technical Report Series*, CRL 97/6, 1997.
9. R. Koch, M. Pollefeys and L. Van Gool. Multi Viewpoint Stereo from Uncalibrated Video Sequences. In H. Burkhardt, B. Neumann, eds., *Computer Vision - ECCV'98, Vol.I*, 55-71, 1998.
10. D. Marr and T. Poggio. A cooperative stereo algorithm. *Science*, 194, 1976.
11. P. S. Penev, J. J. Atick. Local Feature Analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477-500, 1996.
12. E. Trucco, A. Verri. Introductory Techniques for 3-D Computer Vision. Prentice Hall, Upper Saddle River, 1998.
13. T. Vetter and V. Blanz. Estimating coloured 3D face models from single images: An example based approach. In H. Burkhardt, B. Neumann, eds., *Computer Vision - ECCV'98, Vol.II*, 499-513, 1998.