Multi-view passive 3D face acquisition device

L.J. Spreeuwers
Signals and Systems Group, Faculty of EEMCS
University of Twente
P.O. Box 217, 7500 AE Enschede, Netherlands
l.j.spreeuwers@utwente.nl

Abstract: Approaches to acquisition of 3D facial data include laser scanners, structured light devices and (passive) stereo vision. The laser scanner and structured light methods allow accurate reconstruction of the 3D surface but strong light is projected on the faces of subjects. Passive stereo vision based approaches do not require strong light to be projected, however, it is hard to obtain comparable accuracy and robustness of the surface reconstruction. In this paper a passive multiple view approach using 5 cameras in a '+' configuration is proposed that significantly increases robustness and accuracy relative to traditional stereo vision approaches. The normalised cross correlations of all 5 views are combined using direct projection of points instead of the traditionally used rectified images. Also, errors caused by different perspective deformation of the surface in the different views are reduced by using an iterative reconstruction technique where the depth estimation of the previous iteration is used to warp the windows of the normalised cross correlation for the different views.

1 Introduction

Face recognition has been a very active research field for the last 15 years since Turk and Pentland published their face recognition using PCA in 1993 and impressive progress has been achieved. However, illumination, pose and aging remain serious problems. Recently, 3D face recognition has begun to develop and promises improved robustness against illumination and pose variations. Devices used for 3D face acquisition are mostly laser scanners like the Minolta 910, which was used to acquire the 3D images in the FRGC v2 database, or structured light scanners like the viSense from Polygon. Both are active devices that use a controlled light source and produce dense reliable 3D measurements with high depth accuracy. The disadvantages of these active devices are:

- They project light on the subjects face, which may be an unpleasant experience and is bound to strict regulations to avoid damage to the eyes
- Using projection of light and light patterns does not work well if there are strong external light sources, like e.g. the sun
- Theyare not really fit to record videos, because they have a time lag between the first and last scanned point in case of the laser scanner and between successively recorded images with different patterns for the structured light approach.

In principle a passive sensor based on stereo vision techniques does not suffer or suffers less from these defects. No strong light needs to be projected on the face and images can be captured at video speed and there is no lag between different phases of the acquisition. However, classical stereo vision [BBH03] is not really fit for the high resolution images required for 3D face recognition. In this paper, a multiple view approach is presented that enables us to reconstruct 3D facial surfaces with high accuracy. Most notably, the accuracy and robustness are increased by combining measurements from multiple views and allowing perspective deformation of correlation windows using an iterative reconstruction approach.

In the next section first the basics of 3D reconstruction from stereo vision are briefly explained. Next the most important problems entailed with classical stereo vision are described and subsequently the expected improvements using the multi-view approach. Then the iterative reconstruction approach is presented, followed by experiments and results and, finally, conclusions.

2 Basics of 3D reconstruction from stereo cameras

A stereo camera setup consists of 2 cameras. A point P in the 3D world is projected on both left (Q_l) and right images (Q_r). If we know the geometric parameters of the cameras, the coordinates of P can be reconstructed from the image-coordinates of Q_l and Q_r using triangulation. This means we have to find corresponding points in the left and right images. Instead of searching the whole images for corresponding points, the epipolar constraint can be applied: given Q_l, all possible candidates for Q_r are on a line in the right image. This is illustrated in fig.1a.

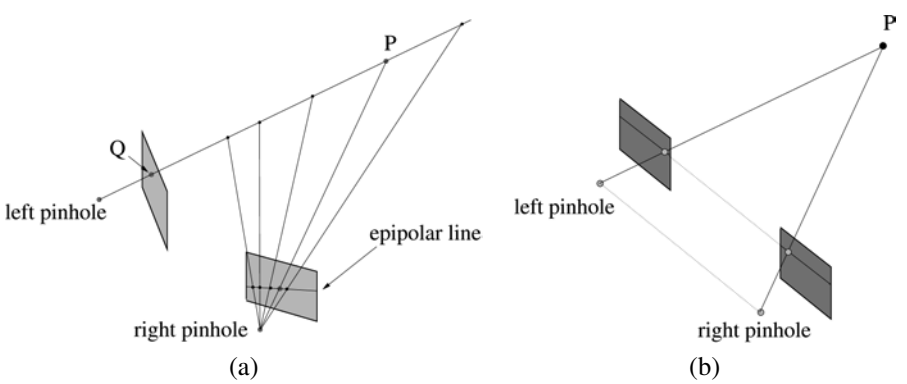


Figure 1: a) Stereo camera, projection of point P on image planes and epipolar line; b) Non-verged camera configuration: image planes are parallel to baseline and all epipolar lines are horizontal lines.

The images can be warped such that for each point in the left image only points on the same row in the right image have to be considered. The warped images are called rectified images and the corresponding camera configuration is called the non-verged stereo camera

configuration (see fig.1b). For a non-verged camera system the image planes of the two cameras are parallel to the baseline of the cameras, i.e. the line between the pin-holes. As a result, all epipolar lines are horizontal lines.

Corresponding points are often found by correlating local neighbourhoods in the rectified images and selecting the points with the highest correlation. The distance between the matching points in the two images is called the disparity and the depth is inverse proportional to this disparity. An example of the normalised correlation as a function of the disparity is given in fig.2.

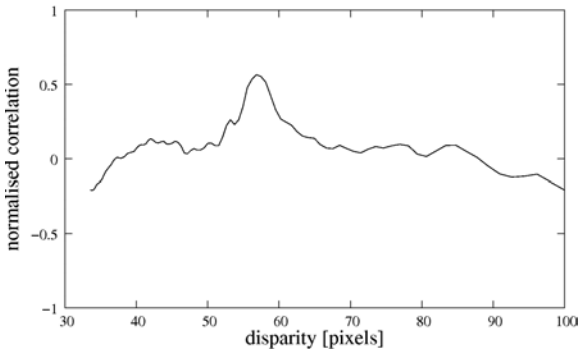


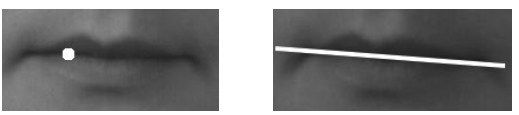
Figure 2: Normalised correlation as a function of the disparity

An in-depth discussion of classical stereo approaches can be found in [BBH03].

3 Typical problems of classical stereo vision

Areas with little texture In areas with little texture, the correlation based methods give similar correlation values for all points. Thus corresponding points cannot be located accurately and, hence the 3d points cannot be reconstructed reliably.

Structures parallel to baseline For structures parallel to the baseline of the cameras (i.e. the line between the pinholes) the correlation between neighbouring positions is more or less the same. This again results in inaccuracy of the reconstructed 3d coordinates. Typical areas in a face that are parallel to the baseline are the mouth (see fig.3 and eyebrows for a horizontally aligned pair of cameras and the nose for vertically aligned cameras.



left image right image + epipolar line

Figure 3: Left and right image of an example of a structure parallel to the baseline: the mouth. The line in the right image is the epipolar line corresponding with the dot in the left image.

Repeating texture Repeating or similar textures along the epipolar lines result in multiple maxima for the correlation and no clear best match, which in turn results in possible incorrect reconstruction of 3d points.

Accuracy of z-measurements The accuracy of the reconstructed 3D coordinates depends on the distance of the 3D object to the cameras. The accuracy in x- and y-directions is inverse proportional to this distance. The accuracy in the z-direction is inverse proportional to the square of this distance. For a stereo setup for acquisition faces at a distance of about 60 [cm], the accuracy in z-direction is 10 times lower than in x- and y-directions.

Occlusions Because both cameras look at the object from different positions, there can be areas on a curved object that are obscured by other parts of the object. In faces, these include the areas around the nose, below the chin and around the ears. The obscured areas are either invisible in both images or only visible in one of the two images. In both cases it is impossible to determine the corresponding 3D coordinates of points in these areas.

Specular reflections The problem with specular reflections is that the position in the images depends on the viewing direction of the cameras. If the camera moves, the position of the specular reflection spot appears to wander over the object. If the specular reflection spots are used for stereo matching, incorrect 3D positions result.

Different perspective projections of surface patches Surface patches that are not parallel to the image planes are projected differently on the different image planes, resulting in possibly incorrect maxima for the correlation and incorrect or less accurate reconstruction of 3D coordinates. This process is illustrated in fig.4: a square neighbourhood of a pixel in the left image is projected on the surface of the object. Because the different points of the surface patch have different depths, the corresponding neighbourhood in the right image is not square anymore. If still a square window is used in the right image to calculate the correlation, we may not find the maximum correlation at the correct position. For curved surfaces, the deformation is even more serious.

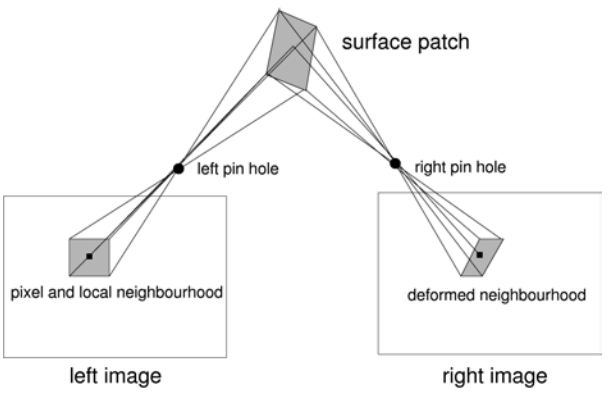
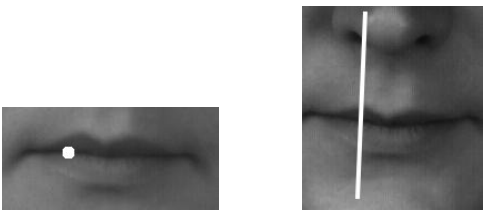


Figure 4: A surface patch is projected with a different perspective deformation on the left and right image planes.

4 Advantages of multi-view approach

Areas with little texture In areas with little texture, having available more measurements helps to improve robustness and accuracy of the 3d reconstruction.

Structures parallel to the baseline If we have available multiple cameras, they can be configured in such a way that there is always a pair of which the baseline is not parallel to the structures in the images. Here, we chose 5 cameras in a '+' configuration which contains horizontal as well as vertical baselines. In fig.5 the epipolar line for a pair of cameras with a vertical baseline is shown. Clearly, along this epipolar line, the texture varies significantly, thus accurate location of matching points is possible.



left image right image + epipolar line

Figure 5: Epipolar line for cameras with vertical baseline

Repeating texture Repeating structures along the epipolar lines can be very efficiently suppressed by using a multiple view approach, because for each view the epipolar lines have a different direction (provided not all cameras are in a row). It is highly unlikely that the repetition of the texture appears in all of those directions.

Accuracy in z-direction By combining multiple independent measurements, the errors tend to cancel out. It has been shown that, theoretically, the uncertainty in the x- and y-directions decreases with $1/\sqrt{K}$, with K the number of cameras, and the uncertainty of the 3D z-coordinate decreases with $1/\sqrt{K^3}$ [För98].

Occlusions Having available more views increases the likelihood that an area is visible in at least two camera views, which enables us to reconstruct the 3d coordinates.

Specular reflections If more views are available, then it is more likely that always at least two images can be selected which do not show specular reflections at a certain position thus allowing us to reconstruct the correct 3d coordinates.

5 3d reconstruction without using rectified images

Since it is impossible to warp the images of more than 2 cameras such that they are all rectified relative to each other, combining results for correlation based multiple views is not straightforward. Therefore, we took a direct approach without using rectified images. If we want to correlate a point Q_A in image A with points in other images, then for all

points of a local neighbourhood of Q_A and all possible depths we calculate the corresponding points in the other images, get the pixel values by interpolation and perform the correlation. This means we do not find the maximum correlation as a function of the disparity, but directly as a function of the depth. This makes combination of the correlation results straightforward. As an extra bonus, this approach allows us to tackle the problem of the perspective deformation of the correlation windows by using an iterative approach where the previous estimate of the 3D surface is used to modulate the relative depths of the points within the correlation mask. This approach is more computationally expensive than the process with rectified images, but by using a clever caching approach, the difference can be minimised. This caching means pre-calculation of the interpolated gray levels for corresponding points for all possible depths in a certain range with a certain step size in depth. Calculation of this depth-gray level cache takes more time than calculation of the rectified images, but, generally, still far less time than the normalised correlation process and thus 3D reconstruction is not so much slower than the traditional approach using rectified images. The construction of the cache for a single pixel for two images is illustrated in fig.6. As can be seen in the figure, for each possible depth of pixel Q in the left image, the corresponding position in the right image is calculated, the gray level is obtained by interpolation and stored in the cache. Normalised correlation is then calculated for all possible depths within the set range.

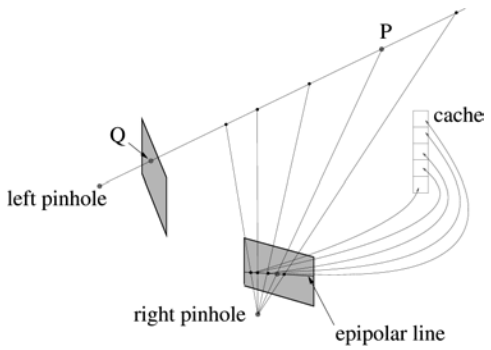


Figure 6: Construction of the depth-gray level cache for 1 pixel and 2 images

6 Experiments and results

6.1 The acquisition system

The system we designed consists of 5 high resolution (1280x960) video cameras that connect to a PC through fire-wire connections. The cameras are synchronised externally to allow them to capture images simultaneously. The cameras are configured in a '+' configuration with one centre camera, one left, one right one upper and one lower camera. By choosing the '+' configuration, there are no structures that are parallel to the baselines of

all camera pairs simultaneously. A photograph of the 5 camera system and the setup for recording a set of images of a subject are shown in fig.7.

Figure 7: 5 camera system and setup for recording a set of images of a subject

Calibration is performed on pairs of cameras relative to the centre camera. The calibration toolbox of OpenCV [Ope] is used, which is a quite advanced and accurate calibration approach using multiple recordings of a chessboard calibration object.The calibration toolbox is mainly based on [Zha99].

6.2 Experiments

In this section, a number of experiments are proposed to show the proposed multi-view approach performs as we expect. The following experiments are presented:

1. Depth estimation of a single point: stereo vs. multi-view for a point in the mouth area (texture parallel to baseline for stereo)
2. Depth estimation of a single point: stereo vs. multi-view for a point in an area with little texture (forehead)
3. Depth estimation of a single point: stereo vs. multi-view for a point on a strongly curved surface (bridge of the nose)
4. Depth estimation of a horizontal line iterative vs. non-iterative approach

For all of these experiments, except for experiment 3, the size of the correlation window was set to 21x21 pixels, because this proved a reasonable compromise between robustness of correlation and accuracy.

Experiment 1 A single point in the central image is picked and the correlation as a function of depth is shown for all 4 camera pair combinations with the central camera. The selected point is on the mouth (see fig.3) and for the horizontal camera pairs we expect a

large range in the depth where the normalised correlation is more or less constant, while for the camera pairs with vertical baseline we expect a relatively sharp peak. In fig.8 the normalised correlation as a function of depth is shown for the described point for four different camera pairs, two with horizontal and two with vertical baselines and a combination of the four correlations by simple averaging.

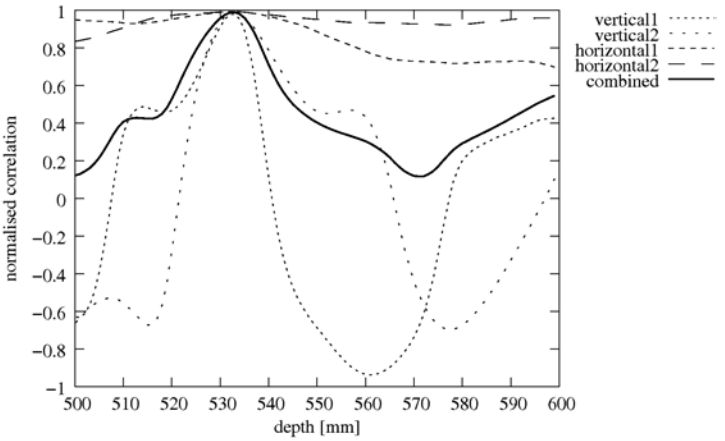


Figure 8: Normalised correlation as a function of depth for four camera pairs and a combination for a point on the mouth

Clearly, the two cameras with horizontal baselines show a relatively flat and high correlation near the actual depth of 533 [mm]. Just as expected it is difficult to obtain an accurate depth estimate using just those cameras. The camera pairs with vertical baseline allow good recovery of the depth. Combination of all four curves by averaging gives a nice peak as well. Because small deviations are are averaged out, we may expect that this curve allows more accurate recovery of the depth than the individual curves.

Experiment 2 Again a single point in the central image is picked and the correlation as a function of depth is shown for all 4 camera pair combinations with the central camera. The selected point is on the forehead, an area with little distinguishing texture and for all camera pairs we expect a large range in the depth where the normalised correlation is more or less constant. In fig.9 the normalised correlation as a function of depth is shown for the described point for four different camera pairs, two with horizontal and two with vertical baselines and a combination of the four correlations by simple averaging.

From the individual curves it is difficult to reliably determine the correct depth. From the combined curve, a unique maximum of the normalised correlation can be found, however, the curve does not show the desired sharp peak. As expected, depth estimation for areas with little texture and constant illumination improves by using multiple views, but not as much as in the case with the textures parallel to the baseline.

Experiment 3 Again a single point in the central image is picked and the correlation as a function of depth is shown for all 4 camera pair combinations with the central camera. The selected point is on the bridge of the nose, an area with strong curvature and we may expect

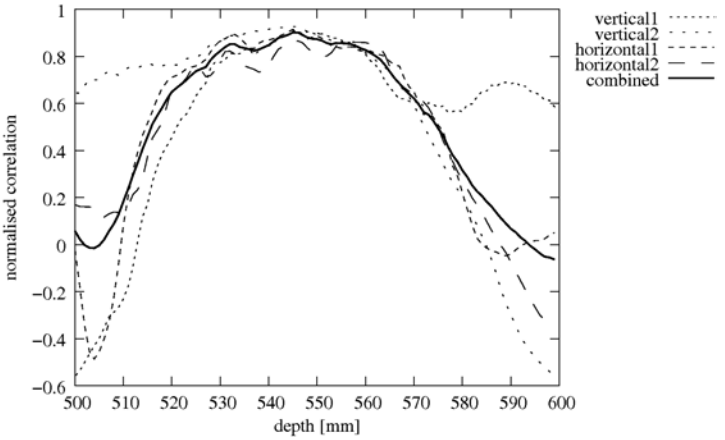


Figure 9: Normalised correlation as a function of depth for four camera pairs and a combination for a point on the forehead

incorrect correlation results due to the deformation of the projections of the correlation window. Since the aim of this experiment is to show how much the correlation can be influenced by this deformation, a synthetic image is used for which the depth values are exactly known. It consists of two ellipsoids: one of the size of a head and one representing the nose. No texture was added, only the lambertian reflection is taken into account.

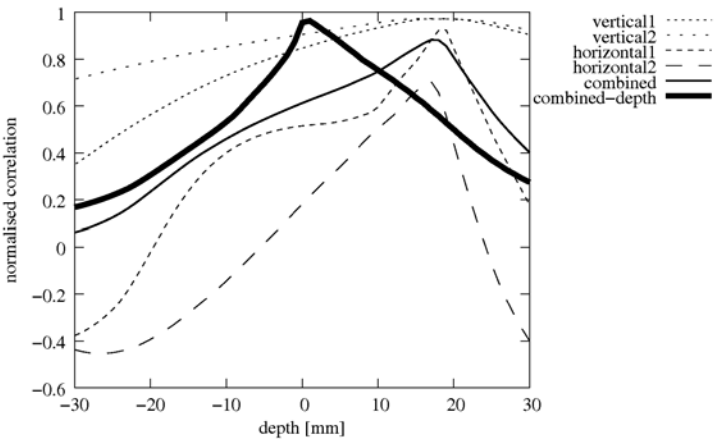


Figure 10: Normalised correlation as a function of depth for four camera pairs and a combination for a point on the bridge of the nose

In fig.10 the normalised correlation as a function of depth is plotted for the described point for four different camera pairs, two with horizontal and two with vertical baselines and a combination of the four correlations by simple averaging. In addition to these curves, a fifth curve (combined-depth) gives the combination of the 4 curves, if the shape-corrected

correlation windows are taken into account using the true depth map of the face. Instead of the absolute depth, the depth relative to the true depth is given on the horizontal axis. As can be seen, the maxima of the normalised correlations show a significant offset. This is the effect that can be expected due to deformation of the correlation windows. If the corrected correlation windows are used, the maximum is on the correct position (combined-depth curve). Incidentally, since the nose is primarily a vertical structure, it is to be expected that the camera pairs with horizontal baselines give a more pronounced peak, which is indeed the case.

Experiment 4 In this experiment the synthetic face model is used, but a piece of a horizontal line across the nose is reconstructed. This will better illustrate the correction resulting from the iterative reconstruction approach. Figure 11 shows the true depth along the line and the estimated depth without applying the iterative approach (iteration 0) and the iterations 1 and 4 where for each iteration the depth map of the previous iteration is used to modulate the shape of the correlation window. Finally, the reconstructed depth when using the true depth to shape the correlation windows is plotted in the "using true depth" curve.

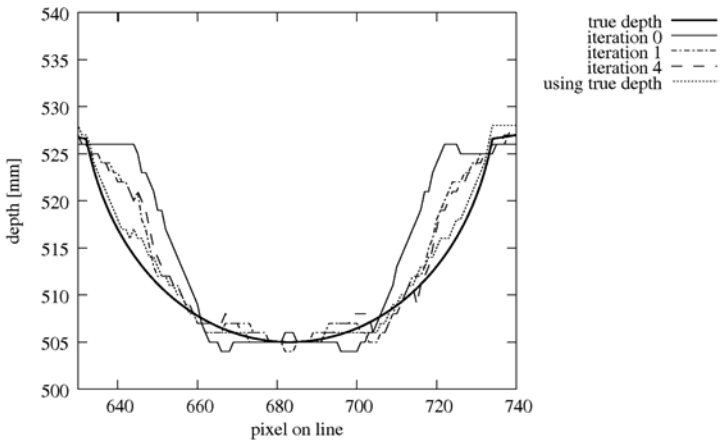


Figure 11: Depth reconstruction along a horizontal line across the nose using the iterative approach

From figure 11 it is clear that the normal reconstruction (iteration 0 curve) gives serious deviations from the ground truth. The iterative approach (iteration 1 and 4 curves) results in a significantly better approximation of the ground truth, although not as good as the reconstruction that applies the true surface to warp the correlation windows, which approximates the true depth very well. The difference between the 4th and the 1st iteration is rather small. We found that normally the first iteration gives the largest improvement.

6.3 Reconstruction of a full face

The method was applied to obtain a reconstruction of a complete face. In figure 12, the recorded images obtained with the 5 cameras in a '+' configuration are shown.

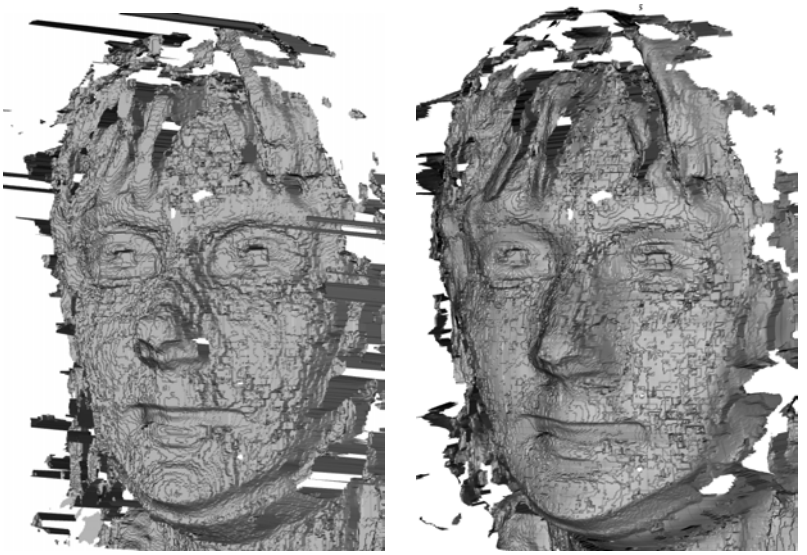
Figure 12: Images obtained using the 5 camera '+' configuration

In figure 13a a reconstruction is shown using the 5 views, while in figure 13b 3 iterations were performed where each time the surface of the previous iteration was used to modulate the depths of the correlation windows. Clearly the latter gives a smoother and more accurate result as already was shown in the previous section with reconstructions of single lines. Note that the ridges near the bridge have gone and that the nose has a different (and more correct) shape in figure 13b. The contours that are still visible are mostly resulting from the fact that for computational efficiency, the face was divided in patches of about 9x9 [mm].

The complete reconstruction of a face to a resolution of 0.5 [mm] in x-, y- and z-direction takes approximately 8 minutes on a single core Pentium 4 at 2.8 GHz.

7 Conclusion

We show that using a multi-view approach to 3d face reconstruction overcomes many of the typical problems of classical stereo reconstruction. The multi-view approach significantly improves robustness and accuracy of the reconstruction in areas with little texture and where in the stereo-view approach structures are parallel to the baseline of the cameras. Furthermore, we propose an approach to 3d reconstruction that does not use rectified images, but finds corresponding points in the other images for all depths. This approach allows combination of the results of multiple views more easily and also allows taking into account the different shapes of correlation windows in the different images using an iterative approach where the surface estimation of the previous iteration is used to "shape" the correlation windows which results in a significant reduction of the error of the depth estimation. The proposed methods have been implemented using a depth-gray level cache which allows for relatively efficient calculation. Several experiments show that the proposed approach results in the expected improvements relative to classical stereo.



(a) (b)

Figure 13: Reconstructions of a complete face; a) single iteration, b) after 3 iterations

References

[BBH03] Myron Z. Brown, Darius Burschka, and Gregory D. Hager. Advances in Computational Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003.

[För98] W. Förstner. On the Theoretical Accuracy of Multi Image Matching, Restoration and Triangulation. In *Festschrift zum 65. Geburtstag von Prof. Dr.-Ing. mult. G. Konecny*. Institut für Photogrammetrie, Universität Hannover, 1998.

[Ope] Open Computer Vision Library. http://sourceforge.net/projects/opencvlibrary.

[Zha99] Zhengyou Zhang. Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In *Proceeedings of the International Conference on Computer Vision*, pages 666–673, Kerkya, Corfu, Greece, 20-25 September 1999. IEEE Computer Society.