Direct Surface Fitting

Nils Einecke, Sven Rebhan, Volker Willert, Julian Eggert

2010

Preprint:

This is an accepted article published in Proceedings of VISAPP 2010 International Conference on Computer Vision Theory and Applications. The final authenticated version is available online at: https://doi.org/[DOI not available]

DIRECT SURFACE FITTING

Nils Einecke, Sven Rebhan, Julian Eggert
Honda Research Institute Europe, Carl-Legien-Strasse 30, 63073 Offenbach, Germany
nils.einecke@honda-ri.de, sven.rebhan@honda-ri.de, julian.eggert@honda-ri.de

Volker Willert
Control Theory and Robotics Lab, TU Darmstadt, Landgraf-Georg-Strasse 4, 64283 Darmstadt, Germany
volker.willert@rtr.tu-darmstadt.de

Keywords: Stereo, Model fitting, Surface estimation, 3-D perception.

Abstract: In this paper, we propose a new method for estimating the shape of a surface from visual input. Assuming a parametric model of a surface, the parameters best explaining the perspective changes of the surface between different views are estimated. This is in contrast to the usual approach of fitting a model into a 3-D point cloud, generated by some previously calculated local correspondence matching method. The main ingredients of our approach are formulas for a perspective mapping of parametric 3-D surface models between different camera views. Model parameters are estimated using the Hooke-Jeeves optimization method, which works without the derivative of the objective function. We demonstrate our approach with models of a plane, a sphere and a cylinder and show that the parameters are accurately estimated.

1 INTRODUCTION

A basic step of many stereo algorithms is the computation of a disparity or depth map by means of a local correspondence search. Instead of comparing single pixels a local window around each pixel is used because pixel comparisons are prone to produce false correspondences. This constitutes a local smoothness assumption, which dramatically improves the detected correspondences. However, some correspondences are still wrong due to repetitive patterns, camera noise or slight view changes between different camera images. In order to remove such erroneous correspondences and to improve the accuracy, more global smoothness assumptions are applied to the resulting disparity maps. A common way of doing so, is to fit basic surface models, e.g. planes (Bleyer and Gelautz, 2005; Hirschmüller, 2006; Klaus et al., 2006; Wang and Zheng, 2008), into the 3-D point data that can be extracted from the disparity maps.

In this paper, we present an alternative approach which integrates parametric surface models directly into the correspondence search. This means that we fit surface models directly to the image data and not into some preprocessed disparity maps. This leads to a much higher accuracy because the original stereo input images carry the complete visual information while the disparity maps contain only the extracted depth information. Furthermore, the model-based correspondence search allows to estimate the depth for large image regions at once, which also improves robustness and accuracy. The basic idea of our approach is to estimate depth by means of the perspective view changes a surface undergoes between different camera views. To achieve this, we describe the perspective view changes of a surface via its parametric description, e.g. center and radius of a sphere or anchor point and rotation angles of a plane. The parameters of a surface model are estimated using Hooke-Jeeves (Hooke and Jeeves, 1961) optimization, which is a direct search method. Its objective is to find those parameters which explain the perspective view changes best.

Early work on incorporating models of the 3-D scene geometry directly into the correspondence search was done by Cernuschi-Frias et al. (Cernuschi-Frias et al., 1989). The authors presented a framework for estimating parameters of different surface models. Although the approach was analyzed in detail on a theoretical level, only a few experimental results were presented. Furthermore, the framework uses an approximation of the pinhole camera model. In contrast, more recent approaches (Baker et al., 1998; Okutomi et al., 2002; Habbecke and Kobbelt, 2005)

are usually using the concept of homography mapping (Hartley and Zisserman, 2004), which does not require such an approximation. For example Habbecke and Kobbelt (Habbecke and Kobbelt, 2005; Habbecke and Kobbelt, 2007) elaborated on this idea by following an approach similar to that of Lucas and Kanade (Lucas and Kanade, 1981). They derived a Gauss-Newton style matching and approximated the partial image derivatives with a first-order Taylor expansion. This leads to an efficient iterative optimization scheme based on image gradients at different resolution scales. Although the results were impressive, their approach has two major limitations. First, the homography transformation limits the approach to planar fitting. Second, the Gauss-Newton optimization is restricted to a sum of squared values, i.e. the objective function cannot be changed. Our approach overcomes these limitations as we use a direct search method (Hooke and Jeeves, 1961) instead of a classical optimization method based on derivatives. In doing so, our approach does not constrain the formulas that describe the perspective view changes of a model, e.g. they can be non-linear and do not need to be differentiable. By this, we go beyond the planar limit and allow for various 3-D models. This also allows for a wide range of objective functions, even non-linear ones like the Sum of Absolute Differences (SAD) or truncated measures.

The paper is organized as follows. In section 2, we sketch a general way of deriving formulas which describe the perspective view changes of a parametric 3-D model. We derive and present the mapping formulas for a plane, a sphere and a cylinder. Section 3 explains our model fitting and parameter estimation method in detail. In section 4, we show that our approach is able to accurately fit different surface models directly to image data. Furthermore, we present a tentative idea of model selection by showing that the most suitable model is the one with the smallest residual error.

2 MATHEMATICAL BASICS

In the following, we derive formulas for transforming surface views from one camera to another, based on a parametric description of a surface (3-D model) and the pinhole camera model. In case of a planar model such a transformation is well-known as homography (Hartley and Zisserman, 2004). Here, the formulas are derived in a different way to motivate the research and usage of other surface models than planes, which the homography is restricted to. In order to make the formulation easier to understand, we assume a parallel camera setting. However, the approach itself is not constrained to such a setting.

2.1 Perspective Projection

In this paper, we consider a rectified, parallel stereo camera setting where the two cameras have the same focal length f (just for convenience). Furthermore, we have two coordinate systems with the origins in the foci of the two cameras. In the following, variables are indexed with L or R to denote whether they belong to the left (L) or right (R) coordinate system. The perspective projections for 3-D points $\mathbf{x} = (x,y,z)^T$ onto the camera CCD chips are

$$\mathbf{u}_L = \frac{f}{z_L}\begin{pmatrix} x_L \\ y_L \end{pmatrix} \quad (1)$$

$$\mathbf{u}_R = \frac{f}{z_R}\begin{pmatrix} x_R \\ y_R \end{pmatrix}, \quad (2)$$

where $\mathbf{u}_L$ and $\mathbf{u}_R$ are the perspective projections of $\mathbf{x}_L$ and $\mathbf{x}_R$, respectively. Note that $\mathbf{u}_L$ and $\mathbf{u}_R$ are two-dimensional chip coordinates with $\mathbf{u} = (u_x, u_y)$. In a parallel stereo system, coordinates of the left coordinate system can easily be transformed into coordinates of the right coordinate system by subtracting the baseline b. Hence the projection equation (2) of the right camera can be rewritten as

$$\mathbf{u}_R = \frac{f}{z_L}\begin{pmatrix} x_L - b \\ y_L \end{pmatrix}. \quad (3)$$

For a correspondence pair $(\mathbf{u}_L, \mathbf{u}_R)$ the 3-D coordinates $\mathbf{x}_L$ of the corresponding 3-D world point can be calculated. The other way around, if the depth of a point is known, it can be mapped from one view to the other. By rearranging the projection equation (1) of the left camera we get

$$x_L = \frac{u_{Lx} \cdot z_L}{f} \quad (4)$$

$$y_L = \frac{u_{Ly} \cdot z_L}{f}. \quad (5)$$

Substituting x_L and y_L into the modified projection equation (3) for the right camera leads to the basic mapping equation

$$\mathbf{u}_R = \mathbf{u}_L - b\frac{f}{z_L}\begin{pmatrix} 1 \\ 0 \end{pmatrix}. \quad (6)$$

By means of the above equation a pixel from the left camera can be mapped to a pixel in the right camera using the known depth z_L. For cameras that are not parallel this equation has to be extended by the relative translation and rotation of two cameras. In order to map a parametric surface, z_L has to be described in terms of the surface's parametric description. In the

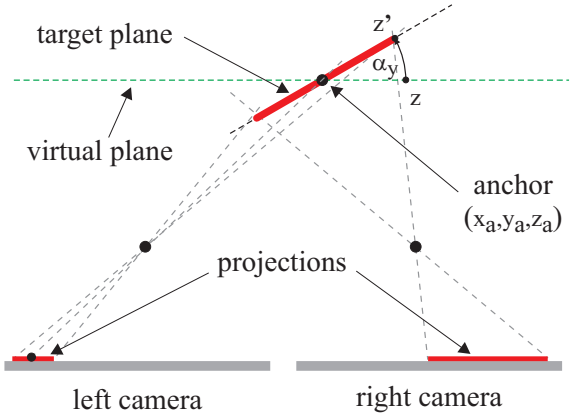


Figure 1: This image shows a schematic configuration of a parallel stereo camera setting and a planar surface, 2-D top view only.

following, we will sketch the derivations for planes, spheres and cylinders. However, the method is applicable in an analogous way to other parametric surfaces.

2.2 Planar Model

In order to derive a formula for z_L that depends on planar model parameters, we describe a planar image region (*target plane*) relative to a *virtual plane* parallel to the CCD-chip. The planes differ by a rotation at a certain anchor point about the x- and y-axis. Figure 1 shows a schematic top view. The anchor point is specified in world coordinates and denoted with $\mathbf{x}_a$. The orientation is specified via rotation angles about the x-axis (α_x) and y-axis (α_y). Note that these two rotations suffice to describe any possible plane orientation. From analytical geometry, it can be derived that points $\mathbf{x}'$ from the *virtual plane* are transformed into points $\mathbf{x}$ on the rotated *target plane* by applying the transformation matrix

$$\mathbf{T} = \begin{pmatrix} \cos\alpha_y & \sin\alpha_x \sin\alpha_y & \cos\alpha_x \sin\alpha_y \\ 0 & \cos\alpha_x & -\sin\alpha_x \\ -\sin\alpha_y & \sin\alpha_x \cos\alpha_y & \cos\alpha_x \cos\alpha_y \end{pmatrix}, \quad (7)$$

leading to the following transformation formula

$$\mathbf{x} = \mathbf{T}\left[\mathbf{x}' - \mathbf{x}_a\right] + \mathbf{x}_a \; . \quad (8)$$

Because the *virtual plane* is parallel to the CCD-chip of the camera, the z-coordinate for points on this frontoparallel plane is always equal to the z-coordinate of the anchor point, $z' = z_a$. Using this, we can rewrite the transformation equation above to

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{T} \begin{pmatrix} x' - x_a \\ y' - y_a \\ 0 \end{pmatrix} + \begin{pmatrix} x_a \\ y_a \\ z_a \end{pmatrix} . \quad (9)$$

With this, the depth z on the *target plane*, given the anchor point and rotation angles, reads as

$$z = (y' - y_a)\sin\alpha_x \cos\alpha_y - (x' - x_a)\sin\alpha_y + z_a, \quad (10)$$

where $(x' - x_a)$ and $(y' - y_a)$ can also be expressed with their counterparts on the rotated *target plane* rearranging and substituting the transformation equations (9):

$$x' - x_a = \frac{x - x_a - (y' - y_a)\sin\alpha_x \sin\alpha_y}{\cos\alpha_y} \quad (11)$$

$$y' - y_a = \frac{y - y_a}{\cos\alpha_x} \; . \quad (12)$$

Applying these two equations to the depth formula (10) and replacing the 3-D world coordinates with their 2-D chip projections (using the projection equations (4) and (5)) finally leads to

$$z_L = f\frac{x_a \sin\alpha_y - y_a \tan\alpha_x + z_a \cos\alpha_y}{u_{Lx}\sin\alpha_y - u_{Ly}\tan\alpha_x + f\cos\alpha_y} \; . \quad (13)$$

With this we have an equation that describes z_L in terms of the parameters of a planar model. Substituting z_L in the basic mapping equation (6) leads to

$$u_{Rx} = u_{Lx} - b\frac{u_{Lx}\sin\alpha_y - u_{Ly}\tan\alpha_x + f\cos\alpha_y}{x_a \sin\alpha_y - y_a \tan\alpha_x + z_a \cos\alpha_y} \quad (14)$$

$$u_{Ry} = u_{Ly} \; . \quad (15)$$

These equations allow for a mapping of the view of a plane from the left camera to the right camera by means of the planar parameters (z_a, α_x and α_y). The values for x_a and y_a can be chosen arbitrarily. They just define at which position the depth z_a of the planar model is estimated. Please note that the mapping equations (14) and (15) for the planar model correspond to the well-known homography transformation. This derivation was done in order to ease the understanding of the derivation of the other models, which are the main focus of this paper.

2.3 Spherical Model

In this section, we show that in our generic framework it is possible to map other parametric surface models starting with the sphere. As in section 2.2, we need to formulate z_L as a function of the parametric model. A sphere in the three-dimensional space with radius r can be described by

$$r^2 = (x - x_a)^2 + (y - y_a)^2 + (z - z_a)^2 \; , \quad (16)$$

where (x_a, y_a, z_a) is the anchor point (center) of the sphere. For a graphical explanation see figure 2. As we have done with the planar equations in section 2.2,

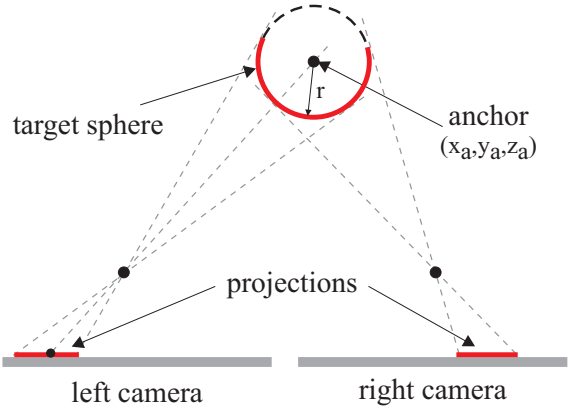


Figure 2: This image shows a schematic configuration of a parallel stereo camera setting and a spherical surface, 2-D top view only.

we replace the 3-D world points with their projections on the CCD-chips using the projection equations (4) and (5). As the replacement is straightforward we omit it for brevity and proceed with the resulting formula rearranged for z_L

$$z_{L1,2} = \frac{\mu \pm \sqrt{\mu^2 - \nu\lambda}}{\lambda} , \tag{17}$$

with

$$\lambda = 1 + \frac{u_{Lx}^2 + u_{Ly}^2}{f^2} \tag{18}$$

$$\mu = z_a + \frac{u_{Lx}x_a + u_{Ly}y_a}{f} \tag{19}$$

$$\nu = x_a^2 + y_a^2 + z_a^2 - r^2 . \tag{20}$$

At a first glance having two solutions in the spherical depth equation (17) looks puzzling. In fact, a closer look at figure 2 reveals that using the "$-$" in the spherical depth equation (17) means mapping a sphere (convex structure) and using the "$+$" means mapping a bowl (concave structure). Therefore, substituting z_L in the basic mapping equation (6) with the spherical depth equation (17) leads to two transformation equations. The first is the equation for transforming the view of a sphere

$$\mathbf{u}_R = \mathbf{u}_L - \frac{bf\lambda}{\mu - \sqrt{\mu^2 - \nu\lambda}} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{21}$$

and the second for transforming the view of a bowl

$$\mathbf{u}_R = \mathbf{u}_L - \frac{bf\lambda}{\mu + \sqrt{\mu^2 - \nu\lambda}} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{22}$$

These equations allow for a mapping of the view of a sphere or a bowl from the left camera to the right camera by means of the spherical model parameters (z_a, x_a, y_a and r).

2.4 Cylindrical Model

The derivation of the formulas for the cylindrical model follows the same scheme like for the planar and spherical model. Since the formulas get a bit lengthy, the following derivation is just a brief sketch. The setup of the cylindrical model is very similar to that of the sphere (see figure 2). We have chosen to describe the cylindrical model by:

$$r^2 = (x - x_a)^2 + (z - z_a)^2 , \tag{23}$$

This means our cylindrical model is infinite in the y-direction. In contrast to the spherical model, it is necessary to incorporate a rotation matrix like we have done for the planar model

$$\mathbf{T} = \begin{pmatrix} \cos\alpha_z & -\sin\alpha_z & 0 \\ \cos\alpha_x \sin\alpha_z & \cos\alpha_x \cos\alpha_z & -\sin\alpha_x \\ \sin\alpha_x \sin\alpha_z & \sin\alpha_x \cos\alpha_z & \cos\alpha_x \end{pmatrix}. \tag{24}$$

For the cylindrical model, we have chosen the rotation about the x-axis and the z-axis. This leads to six parameters for the model of the cylinder with anchor point (a_x, a_y, a_z), rotation angles (α_x, α_z) and radius r. Actually, the model has only five parameters as the y-position for the infinitely expanded cylinder can be fixed. For the derivation we proceed in a way analogous to the plane and the sphere (not shown in full detail here). The resulting depth formula has a structure similar to that of the sphere

$$z_{L1,2} = \frac{\tau \pm \sqrt{\tau^2 - \eta\kappa}}{\kappa} , \tag{25}$$

with

$$\kappa = u_{Lx}^2 \frac{A}{f^2} + u_{Ly}^2 \frac{B}{f^2} + 2u_{Lx}u_{Ly}\frac{C}{f^2} + \frac{2}{f}(u_{Ly}D + u_{Lx}E) + F \tag{26}$$

$$\eta = y_a^2 A + x_a^2 B + 2x_a y_a C + 2z_a(y_a D + x_a E) + z_a^2 F - r^2 \tag{27}$$

$$\tau = u_{Ly}y_a \frac{A}{f} + u_{Lx}x_a \frac{B}{f} + (u_{Ly}x_a + u_{Lx}y_a)\frac{C}{f} + \frac{z_a}{f}(u_{Ly}D + u_{Lx}E) + y_a D + x_a E + z_a F , \tag{28}$$

where

$$A = \sin\alpha_x \sin\alpha_z \cos\alpha_z \tag{29}$$

$$B = 1 - \cos^2\alpha_x \cos^2\alpha_z \tag{30}$$

$$C = \cos^2\alpha_z \tag{31}$$

$$D = 1 - \sin^2\alpha_x \cos^2\alpha_z \tag{32}$$

$$E = -\sin\alpha_x \cos\alpha_x \cos^2\alpha_z \tag{33}$$

$$F = \sin\alpha_x \sin\alpha_z \cos\alpha_z . \tag{34}$$

Substituting z_L of the basic mapping equation (6) with the cylindrical depth equation (25) leads to two transformation equations

$$\mathbf{u}_R = \mathbf{u}_L - \frac{bf\kappa}{\tau \pm \sqrt{\tau^2 - \eta\kappa}} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \qquad (35)$$

These equations allow for a mapping of the view of a cylindrical shape from the left camera to the right camera by means of the cylindrical model parameters (x_a, y_a, z_a, α_x, α_z and r). As was pointed out in section 2.3, "$-$" corresponds to mapping concave structures and "$+$" corresponds to mapping convex structures.

3 MODEL PARAMETER ESTIMATION

The basic idea of our approach is to incorporate models directly into the correspondence search, instead of fitting models into depth or disparity data gained from some local correspondence searches. For this purpose, we derived the transformation equations of the parametric models in the last section that describe the perspective view changes of these models in a stereo camera setting. We now search for the model parameters of larger image regions that explain the perspective view changes of these regions between different camera images. For doing so we use the Hooke-Jeeves (Hooke and Jeeves, 1961) optimization method. Its objective is to minimize the error between the original left view and the transformed right view.

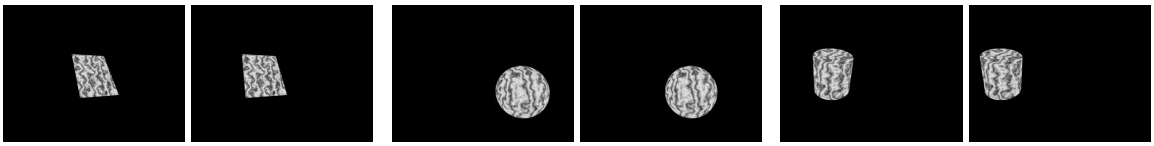
Hooke-Jeeves is a direct search method (Lewis et al., 2000) for optimizing (fitness) functions. Starting from an initial parameter set, an iterative refinement is conducted by sampling alternative parameter sets around the current solution. From these alternative sets the best one is selected. If no better solution is found, the step size is reduced. This is repeated until a minimal step size has been reached. Here we use the SAD between the original left image of a surface and the transformed right image as the fitness function for the Hooke-Jeeves algorithm. This means that the search algorithm tries to find those parameters of a parametric surface that best predict the perspective change between the left and right camera view. We use SAD because it is less sensitive to outliers in the image data compared to a quadratic measure.

It may seem unusual to use Hooke-Jeeves instead of a classical optimization based on gradients. However, direct search methods like Hooke-Jeeves have several advantages over gradient based solutions. First, gradient based approaches need a formal description of the fitness gradient which is based on the image gradients. These, however, can only be approximated locally, e.g. by means of a Taylor expansion (Habbecke and Kobbelt, 2005; Lucas and Kanade, 1981). Because of this, gradient based approaches usually need to rely on a resolution pyramid. There is no such necessity when using a direct search method like Hooke-Jeeves, because it searches the parameter space by means of sampling. Second, it is easy to replace one fitness function with another one, i.e. it is straightforward to exchange the model (transformation formulas) or objective function (matching function). In contrast to this, the formulas in gradient based optimization regimes depend on the model as well as on the used objective function. This means that gradient formulas have to be re-derived when the model or the objective function are changed. Moreover, the possible set of matching metrics is limited, as for example a SAD is not derivable. Last but not least, the Hooke-Jeeves optimization is numerically very stable for the method presented here, since only simple arithmetic and trigonometric functions are used for the transformations.

Notwithstanding its advantages, Hooke-Jeeves is rarely used as it is considered inefficient. Compared to gradient based approaches Hooke-Jeeves needs more iterations. However, the overall speed depends on the function to optimize. Especially, using gradient based approaches on images is quite expensive because for calculating the local gradients the images have to be filtered in each iteration. This filtering is avoided when using a direct search method like Hooke-Jeeves. In (Habbecke and Kobbelt, 2005) a very efficient gradient method for plane estimation was proposed which is about a factor of two to three faster than the Levenberg-Marquardt minimization. Their implementation needs roughly 15 iterations. On an AMD Athlon 64 3500+ they need around 0.2ms for one iteration of a patch of 1000 pixels, i.e. the overall computation time is 3ms. In terms of iterations our Hooke-Jeeves implementation is quite expensive as it usually needs on average 175 iterations. However, on a comparable system (one core of an Intel Xeon X5355) the overall computation time for a patch of 1000 pixels is 6.8ms. This demonstrates that Hooke-Jeeves can compete with state-of-the-art gradient based optimization when it comes to plane fitting.

4 RESULTS

In order to prove the concept of our approach and to evaluate the accuracy of the parameter estimation, we conducted some experiments with virtual scenes.



	Ground Truth	Estimated
α_x	$+37^\circ$	$+37.00^\circ$
α_y	-23°	-22.81°
z_a	500mm	499.70mm

(a) Plane

	Ground Truth	Estimated
x_a	150mm	149.32mm
y_a	-70mm	-69.14mm
z_a	500mm	500.32mm
r	100mm	100.08mm

(b) Sphere

	Ground Truth	Estimated
x_a	-150mm	-150.76mm
y_a	0mm	0.00mm
z_a	500mm	500.21mm
α_x	-31°	-31.62°
α_z	-13°	-11.18°
r	70mm	69.61mm

(c) Cylinder

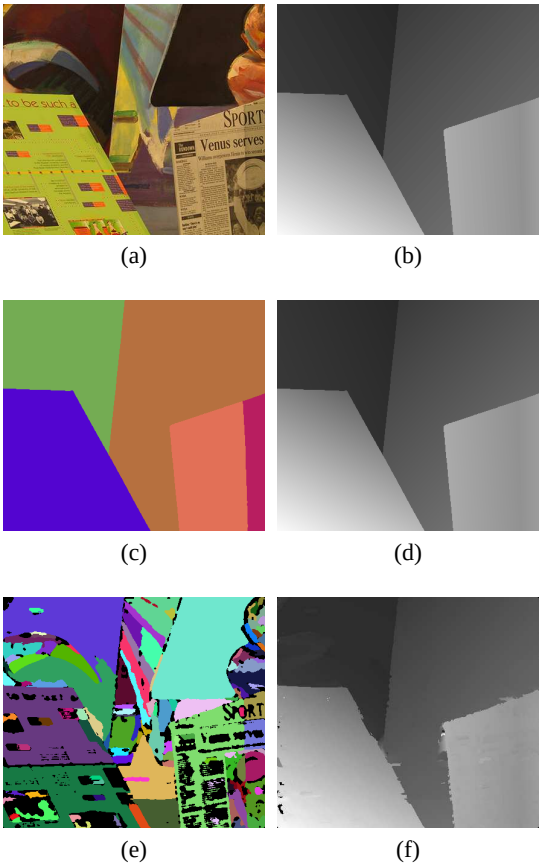
Figure 3: Results of our approach applied on the three different rendered objects a) Plane b) Sphere and c) Cylinder. The images at the top show the left and right camera image of the different objects. The tables below the images show the ground truth parameters of the objects and the parameters estimated with our approach.

To this end, we rendered camera images by means of POVRay (http://www.povray.org/), a free ray-tracing program. We rendered the images such that they corresponded to a standard parallel stereo camera setting. The objects were places in a distance of 50cm in front of the stereo cameras. Figure 3 depicts the rendered images and the results achieved by our approach.

Comparing the ground truth values of the parameters with the estimated parameter values shows that our approach is able to estimate the model parameters very precisely. Although the objects cover only image regions of about 100×100 pixels, angles are estimated up to a half degree for the plane and up to two degrees for the cylinder and positions and radii up to one mm.

In order to evaluate the precision of our approach under more realistic conditions, we used the *Venus* scene from the Middlebury data set (Scharstein and Szeliski, 2003). This scene consists of five planar surfaces. We segment the left image into the five planar regions (figure 4c) in order to estimate planar parameters for each. Note that we segment only the left image, as the search process warps the right image into the left image for comparison. Afterwards we compute a disparity map from the estimated parameters. The results are shown in figure 4. Comparing the ground truth (figure 4b) and the estimated disparity map (figure 4d) reveals almost no errors. The percentage of bad pixels, with an accuracy of 0.5 pixels, is 0.00%, i.e. no erroneous estimations. The percentage of bad pixels is the common error measure used to compare results on the Middlebury data set and is described in (Scharstein and Szeliski, 2003). However, we segmented the image by hand. A standard segmentation algorithm may produce a lot more segments of poorer quality. It is a common assumption in the field of computer vision that homogeneous regions are likely to be planes. Hence, we used a simple region growing algorithm in order to segment the *Venus* scene. Figure 4e shows that such a segmentation leads to a large number of regions of different sizes. Note that regions smaller than 100 pixels are displayed in black. Although this automated preprocessing constitutes quite a challenge for our algorithm, it is still able to produce a good estimation. The percentage of bad pixels (accuracy 0.5 pixels) is 1.39%. This shows that our algorithm is able to estimate model parameters for imperfect and even very small segments as long as the model assumption holds.

For the other models it is much harder to provide a reasonable segmentation. Hence, we investigated if a model selection is possible for a given segment. For this purpose, we had a closer look on what we call the *residual error*. The residual error is the difference between the original left image and the transformed right image, using the parameters estimated by our algorithm. This means that the residual error is the minimal value of the fitness function that has been found by Hooke-Jeeves. However, using the same model the residual error varies substantially for different surfaces. The problem arises mainly from the fact that we use SAD for image comparison. Hence, the residual error tends to be larger for surfaces of high contrast. It has to be analyzed in future work if other objective functions are more suitable. For example using a normalized cross-correlation would make the residual error more descriptive. Because of the variation of the residual error over different surfaces, we decided to compare the residual error of different models. Table 1 shows the residual error of the planar, spherical and cylindrical model applied to the three POVRay



(a) (b) (c) (d) (e) (f)

Figure 4: Results on the *Venus* scene from the Middlebury data set. a) Left camera image, b) ground truth disparity, c) segmentation of the image by hand and e) image segmentation into homogeneous regions using region growing. d) and f) show the disparity maps produced by our approach. Here we applied the planar model to each segmented region and calculated disparity values from the estimated parameters.

rendered objects plane, sphere and cylinder shown in figure 3. The results show a clear difference between the residual error of the correct and wrong models. In most cases the residual error of the wrong models is a magnitude larger than the residual error of the correct model. This means that the correct model can be chosen by taking the one with the smallest residual error. The only exception is the relatively low residual error of the cylindrical model on the plane object. The reason is that the cylindrical model is able to approximate a planar surface well by using a large radius. Although the same argument applies to the spherical model the maximal step sizes used for Hooke-Jeeves restricted such an approximation.

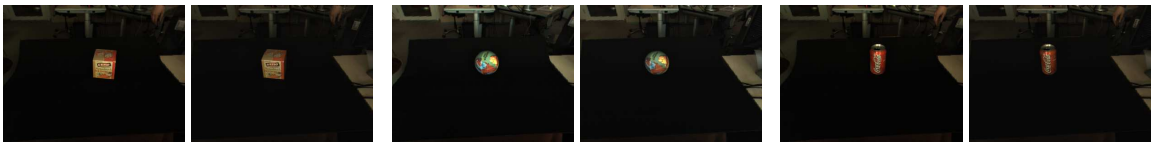
In the last two experiments, we used a real stereo camera system in order to acquire stereo images of real-world objects under real-world conditions. Unfortunately, only partial ground truth data is available here. Figure 5 shows the stereo images of a box, a ball and a can. Below the images of the ball and the can the estimated radius is compared to the radius measured by hand. As you can see the estimation is quite accurate despite of the fact that the objects are really small in size. Comparing the rotation angle α_x of the front face with that of the top face of the box shows that the faces differ approximately by 85°. This is very close to the 90° the faces should differ and is a strong indicator that the estimation was correct. In order to get an impression of how our approach works with imperfect objects and cluttered scenes, we arranged a scene with an apple, a bottle and a box. Figure 6 shows that scene and the estimated disparities of our approach compared to disparities extracted using a standard block matching stereo approach with normalized cross-correlation. For better visibility, we zoomed in the disparity map and removed the background using the object masks. The results show that our approach is able to produce very smooth disparity maps compared to the standard approach. Although the apple and the bottle do not have the exact shape of a sphere and a cylinder our approach is able to fit the models and produce reasonable depth results. Furthermore, matching large regions enhances robustness against clutter in the background and reduces the aperture problem.

Table 1: Comparison of the residual errors (SAD per pixel) of the three different models applied to the three different objects.

	Plane	Sphere	Cylinder
Planar Model	**3.38**	29.73	24.34
Spherical Model	14.45	**5.19**	22.02
Cylindrical Model	7.90	23.44	**6.08**

5 SUMMARY

In this paper, we presented a method which is able to fit 3-D surface models directly in stereo camera images. This is in contrast to the usual approach of fitting models in the disparity data, calculated in advance with a standard stereo method. Prior approaches that fit 3-D surfaces directly to the images are usually restricted with respect to the surface model, camera model or objective function. The major difference in our approach is that we use the Hooke-Jeeves optimization instead of a classical optimization method based on derivatives. This enables literally arbitrary surface models, camera models and objective functions. We demonstrated this by deriving formulas for a planar, a spherical and a cylindrical model. Using rendered scenes, we showed that



	Top Face	Front Face
α_x	$+48.48^\circ$	-36.92°
α_y	$+8.44^\circ$	$+15.11^\circ$
z_a	496.52mm	498.76mm

(a) Box

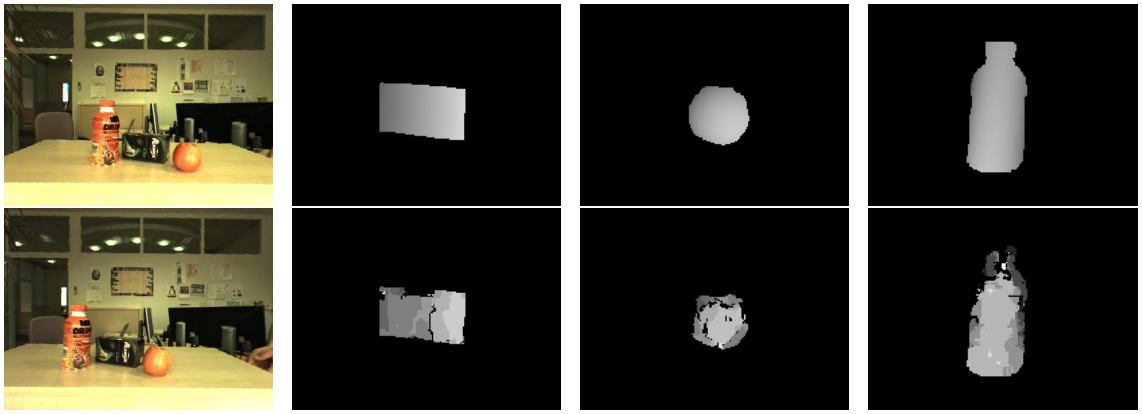
	Ground Truth	Estimated
r	49.30mm	53.98mm
z_a	$\sim$ 525mm	542.19mm

(b) Ball

	Ground Truth	Estimated
r	35.00mm	35.40mm
z_a	$\sim$ 540mm	542.84mm
α_x	—	-23.64°
α_z	—	$+0.30^\circ$

(c) Can

Figure 5: Results of our approach applied to three different real world objects a) Box b) Ball and c) Can. The images at the top show the left and right camera image of the different objects. The tables below the Ball and the Can show the ground truth radius compared to the estimated radius. For the Box the result for the two visible faces are shown, the estimations show that the angle between them is close to 90°.



(a) Office scene (b) Box disparity (c) Apple disparity (d) Bottle disparity

Figure 6: This figure shows the results of our approach compared to a standard stereo approach. a) Top and bottom image show the left and right stereo image, respectively. b-d) Close-ups of the disparities for the three objects Box, Apple and Bottle. The top row shows the disparity maps of our approach and the bottom row the results of a standard block matching stereo approach with normalized cross-correlation.

model parameters are estimated very accurately. Furthermore, we showed that our approach works well under real-world conditions.

In future work, we want to derive formulas for mapping further models, like cones and ellipsoids. With such a set of models available a wide range of applications is conceivable. For example the fitting can be used to generate a coarse pre-classification to aid object recognition. Another important point for future work is to conduct a more elaborated analysis of the accuracy of the parameter estimation and the impact of occlusion. Last but not least, we want to analyze the influence of different objective functions on the robustness and accuracy of the parameter estimation.

REFERENCES

Baker, S., Szeliski, R., and Anandan, P. (1998). A layered approach to stereo reconstruction. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 434–441.

Bleyer, M. and Gelautz, M. (2005). Graph-based surface reconstruction from stereo pairs using image segmentation. In *Videometrics VIII*, volume 5665, pages 288–299.

Cernuschi-Frias, B., Cooper, D. B., Hung, Y.-P., and Belhumeur, P. N. (1989). Toward a model-based bayesian theory for estimating and recognizing parameterized 3-d objects using two or more images taken from different positions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(10):1028–1052.

Habbecke, M. and Kobbelt, L. (2005). Iterative multi-view plane fitting. In *Vision, Modeling, Visualization VMV'06*, pages 73–80.

Habbecke, M. and Kobbelt, L. (2007). A surface-growing approach to multi-view stereo reconstruction. *Computer Vision and Pattern Recognition*, pages 1–8.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition.

Hirschmüller, H. (2006). Stereo vision in structured environments by consistent semi-global matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2386–2393.

Hooke, R. and Jeeves, T. A. (1961). "Direct Search" Solution of Numerical and Statistical Problems. *Journal of the Association for Computing Machinery*, 8(2):212–229.

Klaus, A. S., Sormann, M., and Karner, K. (2006). Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 15–18.

Lewis, R. M., Torczon, V., and Trosset, M. W. (2000). Direct search methods: Then and now. *Journal of Computational and Applied Mathematics*, 124:191–207.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679.

Okutomi, M., Nakano, K., Maruyama, J., and Hara, T. (2002). Robust estimation of planar regions for visual navigation using sequential stereo images. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, pages 3321–3327.

Scharstein, D. and Szeliski, R. (2003). High-accuracy stereo depth maps using structured light. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 195–202, Madison, WI,.

Wang, Z. F. and Zheng, Z. G. (2008). A region based stereo matching algorithm using cooperative optimization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8.