3D Pose Estimation Based on Multiple Monocular Cues

Björn Barrois and Christian Wöhler
DaimlerChrysler Group Research, Environment Perception
P. O. Box 2360, D-89013 Ulm, Germany
{bjoern.barrois, christian.woehler}@daimlerchrysler.com

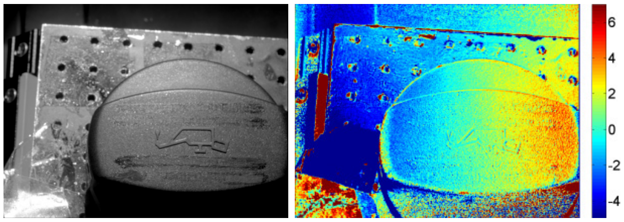
Abstract

In this study we propose an integrated approach to the problem of 3D pose estimation. The main difference to the majority of known methods is the usage of complementary image information, including intensity and polarisation state of the light reflected from the object surface, edge information, and absolute depth values obtained based on a depth from defocus approach. Our method is based on the comparison of the input image to synthetic images generated by an OpenGL-based renderer using model information about the object provided by CAD data. This comparison provides an error term which is minimised by an iterative optimisation algorithm. Although all six degrees of freedom are estimated, our method requires only a monocular camera, circumventing disadvantages of multiocular camera systems such as the need for external camera calibration. Our framework is open for the inclusion of independently acquired depth data. We evaluate our method on a toy example as well as in two realistic scenarios in the domain of industrial quality inspection. Our experiments regarding complex real-world objects located at a distance of about 0.5 m to the camera show that the algorithm achieves typical accuracies of better than 1 degree for the rotation angles, 1–2 image pixels for the lateral translations, and several millimetres or about 1 percent for the object distance.

1. Introduction

3D pose estimation is an important problem in many applications of computer vision and photogrammetry. The problem of pose estimation corresponds to a determination of the rotation and the translation of an object relative to the camera, given the 3D model points and the corresponding 2D perspective projection points in the image. This problem is also known as the exterior orientation problem in the photogrammetric literature [12]. An early survey of pose estimation methods based on the bundle adjustment technique is given in [23]. In the field of computer vision, a first description of the pose estimation problem is given in [8]. The term 2D-3D pose estimation is defined in [10] as an estimation of the pose of a 3D object in 2D input data, for example an intensity image. The geometrical and mathematical problem is regarded in [16], where an edge-based solution is provided. In [15] groupings and structures in the image which are likely to be invariant over a wide range of viewpoints are formed by perceptual organisation. The search space during model based matching is reduced based on a probabilistic ranking method. Another monocular pose estimation approach is described in [20], which exploits point and line correspondences by minimising a suitably chosen error function. The problem of object recognition and localisation is addressed in [18]. The object is represented in a probabilistic framework as a parametric probability density, and the recognition process basically relies on the Bayes rule. The problem of 2D-3D pose estimation of 3D free-form surface models is discussed in [22]. The object is modelled as a two-parametric surface model represented by Fourier descriptors, and the pose estimation problem is solved in the framework of conformal geometric algebra. An edge-based pose estimation approach is described in [24]. In that work, the Chamfer matching technique is used to force convergence of a hierarchical template matching approach. In [17] an object representation based on reflectance ratios is introduced which is used to recognise objects from monocular brightness images of the scene. Pose estimation is performed relying on the reflectance ratio representation and the known geometric object properties. A pose estimation approach that combines intensity and edge information extracted from the input image is described in [19].

Classical monocular pose estimation approaches have in common that they are not able to estimate the distance to the object at reasonable accuracy, since the only available information is the scale of a known object in the resulting image. Scale information yields no accurate results since for small distance variations the object scale does not change significantly. In comparison, for a convergent stereo setup with a baseline similar to the object distance, for geometrical re-



(a) Intensity image. (b) Polarisation angle image.

Figure 1. Example of a high-dynamic range intensity image (grey values are scaled logarithmically) and a polarisation angle image (colour map is scaled in degrees).

sons a depth accuracy of the same order as the lateral translational accuracy is obtainable. For this reason, a variety of 3D pose estimation methods relying on multiple images of the scene have been proposed more recently. For example, a fast tracking algorithm for estimating the pose of an automotive part from a pair of stereo images is presented in [25]. In [21], the iterative closest point algorithm for 3D pose estimation in stereo image pairs is compared with a numerical scheme which is introduced in the context of optical flow estimation. A quantitative evaluation of the two methods and their combination is performed, demonstrating that the highest stability and most favourable convergence behaviour is achieved with the combined approach.

However, many industrial applications of pose estimation methods for quality inspection purposes impose severe constraints on the hardware to be used with respect to robustness and easy maintenance. Hence, it is often not possible to utilise stereo camera systems since they have to be recalibrated regularly, especially when the sensor unit is mounted on an industrial robot. As a consequence, employing a monocular camera system may be favourable from the practical point of view while nevertheless a high pose estimation accuracy is required to detect subtle deviations between the true and the desired object pose.

The pose estimation approach presented in this study exploits the only information in a monocular image apart from scaling which provides an information about the object distance: the amount of defocus. Depth from defocus methods (cf. [2] for a detailed survey) yield a relation between the amount of defocus in the scene and the distance to the camera, allowing to estimate a depth value for each image pixel if texture is present. The accuracy of depth from defocus methods is clearly inferior to that of multi-viewpoint methods such as stereo vision or structure from motion [5] but may provide much more accurate depth cues than merely using scale information.

In the presence of cluttered background or low contrast between object and background, edge information tends to be an unreliable cue for pose estimation. Hence, apart from edge information and defocus, our approach takes into account intensity and polarisation information extracted from the input image data. Such photometric approaches are commonly used for 3D surface reconstruction purposes [3, 11]. Hence, we exploit four complementary sources of radiometric, geometric, and real-aperture information about the scene (intensity, polarisation, edges, and defocus) which we combine in a multi-cue approach to estimate the six degrees of freedom of a rigid object in 3D space. We will evaluate our approach in realistic scenarios related to industrial quality inspection.

2. Combined approach to monocular pose estimation

2.1. Intensity information

A well-known method for 3D surface reconstruction is shape from shading [3, 11]. This approach is based on the so-called reflectance function R_I, which provides the intensity of the light reflected by the object surface depending on the surface orientation, the camera position, and the position of the light source. In the scenarios regarded in this study, we always assume a point light source. In [3] a formulation of the reflectance function of a specular surface is introduced which is based on a diffuse Lambertian component, a broad specular lobe, and a narrow specular spike according to

$$R_I(\theta_i;\theta_r) = \rho\left[\cos\theta_i + \sum_{j=1}^{K}\sigma_j(\cos\theta_r)^{m_j}\right]; \quad (1)$$

where θ_i denotes the incidence angle and θ_r the angle between the viewing direction and the direction of mirror-like reflection. We found experimentally that for the surfaces regarded in our experiments it is appropriate to assume $K = 2$ specular components (cf. Section 3.3). The parameter ρ denotes the surface albedo, which is defined here as a factor depending on the camera lens, the surface reflectivity, the brightness of the light source, and the sensitivity of the camera sensor [11]. It is generally not possible to directly measure this parameter, such that we estimate it in the optimisation algorithm. Although we regard objects of uniform surface albedo in our experiments, our framework would in principle allow to render and investigate objects with a textured surface by using texture mapping in combination with an estimation of the factor ρ. The other parameters of the reflectance function, $\{\sigma_j\}$ and $\{m_j\}$, are determined empirically, regarding a sample of the corresponding surface material attached to a goniometer [3].

We utilise this reflectance function and a CAD model of the object to generate a synthetic image of the observed scene. We implemented an OpenGL-based renderer. Since surface orientation is required for each point of the object surface to compute a reflectance value according to Eq. (1) but OpenGL does not directly provide this information, the

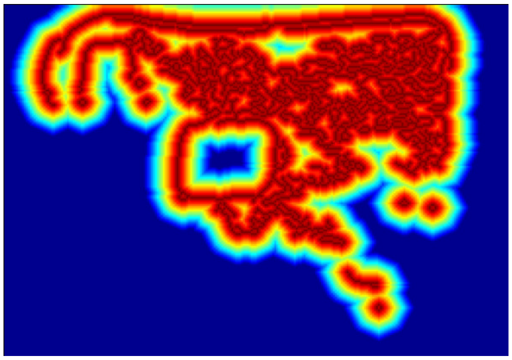


Figure 2. Example of a distance-transformed edge image.

technique developed in [7] is used to calculate the surface normal for every pixel. Afterwards, the reflectance function (1) is used to compute the predicted intensity for each pixel. We obtain a photorealistic image I_S which can be compared with the input image I_I, resulting in the intensity error term

$$e_I = \sum_{u,v} \left[I_I(u,v) - I_S(u,v) \right]^2 ; \qquad (2)$$

where the summation is carried out for the rendered pixels representing the object surface. A disadvantage of the technique proposed in [7] is the fact that no shadow information is generated for the scene. Hence, shadows are computed in a further raytracing step after the photorealistic rendering process.

As the dynamic range of the CCD camera used for our experiments is not sufficiently high to cover both the diffuse and the specular reflectance components, we acquire a series of images of the scene over a wide range of shutter times, combining the individual frames into a single high dynamic range image as described e. g. in [6] (cf. Fig. 1a).

2.2. Edge information

We compute a binarised edge image from the observed intensity image using the well-known Canny edge detector [1]. In a second step, a distance transform image I_D is obtained by computing the Chamfer distance for each pixel [9] (cf. Fig. 2). As our approach compares synthetically generated images with the observed image, we use a modified Chamfer matching technique which is related to the approach described in [24]. We extract the edges in the rendered image with a Sobel edge detector, resulting in a Sobel magnitude image I_E, which is not binarised. To obtain an error term which gives information about the quality of the match, a pixel-wise multiplication of I_D by I_E is performed. The advantage of omitting the binarisation is the continuous behaviour of the dependence of the resulting error function on the pose parameters, which turned out to be a favourable property with respect to the optimisation stage. If the edge image extracted from the rendered image is binarised, the error function becomes discontinuous, making the optimisation task more difficult. Accordingly, the edge error term e_E is defined as

$$e_E = -\sum_{u,v} I_D(u,v) I_E(u,v); \qquad (3)$$

where the summation is carried out over all image pixels (u,v). The minus sign in Eq. (3) arises from the fact that our optimisation scheme aims at a determination of the minimum of the error function.

2.3. Polarisation information

Similar to the intensity, the polarisation angle of the light reflected from the object surface provides information about the rotation of an object relative to the camera. The advantage of using intensity in combination with the polarisation angle is the fact that these quantities contain complementary information about surface orientation [3].

As the scene is illuminated with unpolarised light, the polarisation properties of the reflected light can be measured with a linear polarisation filter mounted in front of the camera lens. When the polarisation filter is rotated around the optical axis, the intensity of each pixel follows a sinusoidal function depending on the orientation angle ω of the filter. We observe the scene at five different orientations of the polarisation filter and fit a function of the form

$$I(\omega) = I_c + I_v \cos[2(\omega - \Phi)] \qquad (4)$$

to the observed pixel intensities. This procedure immediately yields the polarisation angle and the polarisation degree $D = I_v / I_c$.

The behaviour of the polarisation degree tends to vary across the surface in a rather unpredictable, erratic manner. This is especially true for the maximum observed amount of polarisation. Such variations of the polarisation degree are due to its strong dependence on the local microscopic surface roughness. In contrast, the behaviour of the polarisation angle turns out to show a behaviour which is independent of the location on the part surface for the materials regarded in our experiments. The polarisation degree may be a useful and well-defined cue for smooth dielectric surfaces but turned out to be an unreliable feature in the scenarios regarded in this work. Hence, we will utilise intensity and polarisation angle as photometric cues in our pose estimation framework (cf. Fig. 1).

The polarisation angle is favourably described in terms of the surface gradients p and q in horizontal and vertical image direction, respectively, where the coordinate system is chosen such that the scene is illuminated from the right. We now define a reflectance function R_Φ for the polarisation angle, for which we assume an incomplete third-order polynomial of the form

$$R_\Phi(p,q) = a_\Phi + b_\Phi pq + c_\Phi q + d_\Phi p^2 q + g_\Phi q^3 \qquad (5)$$

(cf. [5]). The analytic form of the reflectance function R_Φ is antisymmetric in q as long as the polarisation angle only depends on the azimuth difference between camera and light source but not on the azimuth angles themselves. The parameters a_Φ, b_Φ, c_Φ, d_Φ, and g_Φ depend on the direction to the light source and the viewing direction. They are empirically determined by fitting Eq. (5) to orientation-dependent polarisation data acquired with a goniometer. The renderer is then able to predict the polarisation angle for each pixel. The error term e_Φ for the polarisation angle is defined by

$$e_\Phi = \sum_{u,v} [(u,v) \quad R_\Phi(p(u,v), q(u,v))]^2 ; \qquad (6)$$

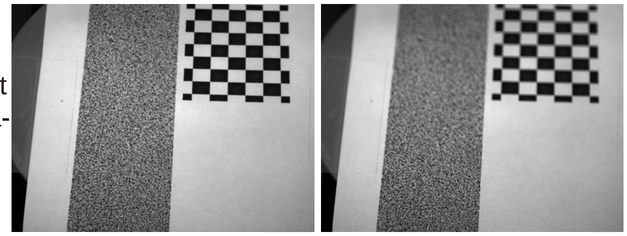
where (u,v) is the polarisation angle observed for pixel (u,v) and $R_\Phi(p(u,v), q(u,v))$ the rendered polarisation angle. Note that for the computation of e_Φ it is necessary to account for the periodicity of the polarisation angle.

2.4. Depth from defocus

A point situated in front of the camera at a distance $z = z_0$ is well focused if z_0 is taken to define a plane on which the camera is focused. By deviating the value of z from z_0 the point appears more and more blurred. This behaviour of real-aperture lens systems is exploited by the depth from defocus approach (cf. [2] for an overview).

An exact description of the point spread function (PSF) due to diffraction of light at a circular aperture is given by the radially symmetric Airy pattern $A(r) \propto [J_1(r)/r]^2$, where $J_1(r)$ is a Bessel function of the first kind. For practical purposes, however, when a variety of additional lens-specific influencing quantities (e. g. chromatic aberration) is involved, the Gaussian function is a reasonable approximation to the PSF [2]. Accordingly, the amplitude spectrum of the Fourier transform of the PSF is also of Gaussian shape displaying a z-dependent width parameter (z) which decreases with increasing amount of defocus.

Basically, we utilise the depth from defocus technique described in [4] to estimate depth values from the amount of defocus. This approach requires two pixel-synchronous images, one of which is acquired with a small aperture, e. g. $f=8$, while the second one is acquired with a large aperture, e. g. $f=2$. This procedure may be automated using a lens equipped with a motorised iris. For the first image we assume that no perceivable amount of defocus is present. The images are partitioned into windows of 32×32 pixels size. After Tukey windowing, the PSF width parameter in frequency space is computed by fitting a Gaussian to the quotient of the amplitude spectra of the corresponding windows of the first and the second image, respectively. Only the range of intermediate spatial frequencies is regarded in order to reduce the influence of noise on the resulting value for . This technique and alternative methods are described in detail in [2].



(a) Sharp input image ($f=8$). (b) Unsharp input image ($f=2$).

Figure 3. Calibration rig for the depth from defocus method.

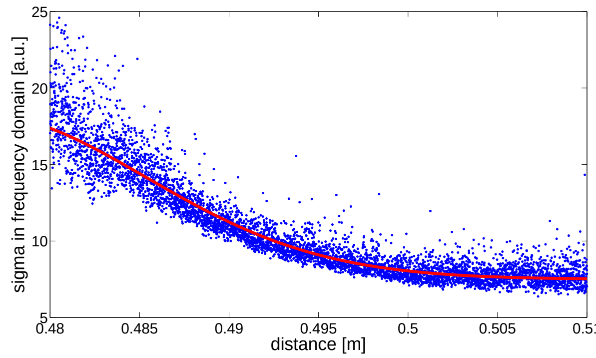


Figure 4. Established relation between depth and defocus.

To calibrate the depth from defocus method we establish the relation between the amount of defocus (z) and the related absolute depth value z. For this purpose we use the calibration rig shown in Fig. 3, which displays on the left a random noise pattern which is especially suitable for estimating the PSF, and on the right a chequerboard pattern of known size to estimate absolute depth values, assuming that the camera is calibrated. Plotting the estimated defocus values (z) over the determined absolute depth values z we obtain the diagram shown in Fig. 4. For the relation between the PSF width parameter (z) in frequency space and the object distance z, in [14] the so-called depth-defocus function

$$\frac{1}{(z)} = \frac{1}{\ _1} e^{-\frac{1}{2}\left(\frac{fz}{z-f} - b\right)^2} + \ _3 \qquad (7)$$

with the parameters $_1$, $_2$, and $_3$ is derived. In Eq. (7) f is the focal length of the camera and b the distance between the lens and the camera sensor determined by internal camera calibration [13]. Eq. (7) is obtained based on the lens law $1/z_0 + 1/b = 1/f$ [14]. The red curve in Fig. 4 shows the result of the fit of Eq. (7) to the measured (z) data points.

2.5. Total error optimisation

To start the optimisation process, an initital object pose has to be provided. With this pose a first set of images (intensity, polarisation angle, edges, and depth map) is rendered. Each measured cue provides an error term, denoted by e_E, e_I, e_Φ, and e_D, respectively. We use these error

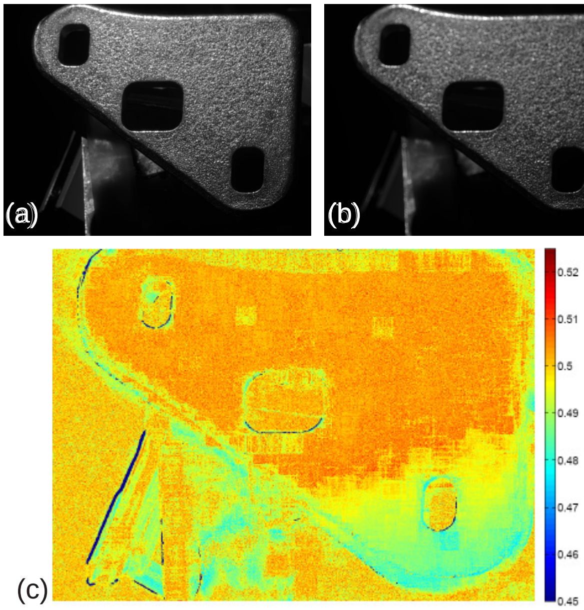


Figure 5. Example of a depth map obtained with the depth from defocus method. (a) Sharp input image, acquired at $f=8$. (b) Unsharp input image, acquired at $f=2$. (c) Resulting depth map. For the black pixels no depth value could be computed. The colour map is scaled in metres.

terms to compute an overall error e_T which is minimised in order to obtain the object pose. As the individual error terms are of different orders of magnitudes, we introduce the weight factors λ_E, λ_I, λ_Φ, and λ_D to appropriately take into account the individual terms in the total error e_T:

$$e_T = \lambda_E e_E + \lambda_I e_I + \lambda_\Phi e_\Phi + \lambda_D e_D. \quad (8)$$

The values of the weight factors are chosen inversely proportional to the typical relative measurement error, respectively.

We found that the influence on the observed intensity, polarisation, edge, and depth cues is different for small variations of each pose parameter (cf. Table 1). For example, a slight lateral translation has a strong influence on the edges in the image but may leave the observed intensity and polarisation angle largely unchanged. On the other hand, under certain viewing conditions, rotations around small angles are hardly visible in the edge image while having a significant effect on the observed intensity or polarisation

	Intensity, polarisation	Edges	Depth
Rotation angles	strong	weak	weak
Lateral translation (x, y)	weak	strong	weak
Translation in z	weak	weak	strong

Table 1. Influence of small changes of the pose parameters on the observed photopolarimetric, geometric, and depth cues.

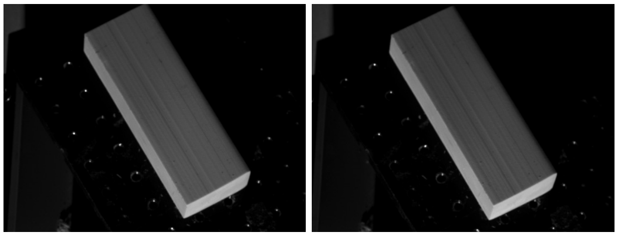
behaviour.

For minimisation of the overall error e_T we use an iterative gradient descent approach. We have chosen this algorithm because of its stable convergence behaviour, but other optimisation methods are possible. Since it is impossible to calculate analytically the derivatives of the total error term with respect to the pose parameters as the error term is computed based on rendered images, the gradient is evaluated numerically. If a certain cue does not provide useful information (which may e. g. be the case for polarisation data when the surface material only weakly polarises the reflected light, or for edges in the presence of cluttered background), this cue can be neglected in the optimisation procedure by setting the corresponding weight factor in Eq. (8) to zero. We will show experimentally in Section 3 that pose estimation remains possible when relying on merely two or three different cues.

Our framework requires a-priori information about the object pose for initialisation of the nonlinear optimisation routine, such that it is especially useful for the purpose of pose refinement. In comparison, the template matching based approach in [24] yields five pose parameters without a-priori knowledge (the distance to the object is assumed to be exactly known). In the addressed application domain of industrial quality inspection, a-priori information about the pose is available from the CAD data of the part itself and the workpiece to which it is attached. Here it is not necessary to detect the part in an arbitrary pose but to measure small differences between the true pose parameters and those desired according to the CAD data. Hence, when applied in the context of industrial quality inspection, our method should be initialised with the pose given by the CAD data, and depending on the tolerances stored in the CAD data, a production fault is indicated when the deviation of one or several pose parameters exceeds the tolerance value. The experimental evaluation described in the next section will show that our framework is able to detect small differences between the true and the desired object pose.

3. Experimental results

To evaluate the performance of the presented approach we estimated the pose of three different test objects and compared the results to the independently derived ground truth. In all experiments, the images were taken with a Baumer industrial CCD camera of 1032×776 pixels image size, equipped with a $f = 25$ mm lens. The approximate distance to the object was 0.5 m. To increase the signal-to-noise ratio of the intensity and polarisation data, the images were downscaled to 258×194 pixels, corresponding to a lateral resolution of about 0.4 mm per pixel. Depth from defocus analysis was performed based on the full-resolution images acquired at apertures of $f=8$ and $f=2$, respectively. The coordinate system was chosen such that the x and y



(a) Input image (pose 1) (b) Input image (pose 2)

Figure 6. Input intensity images for the rubber example.

(a) Input image (pose 1). (b) Input image (pose 2).

Figure 7. Input intensity images for the oil cap example. Greylevels are displayed in logarithmic scale.

axes correspond to the horizontal and vertical image axis, respectively, while the z axis is parallel to the optical axis. The scene was illuminated with a LED point light source located at a known position. For each configuration, the algorithm was initialised with four poses, differing by several degrees in the rotation angles and a few millimetres in translation. As the result of pose estimation we adopted the minimisation run yielding the lowest residual error according to Eq. (8).

3.1. Rubber (toy example)

For our first test we have chosen an object with a simple geometry, a cuboid-shaped rubber. The reflectance function R_I was determined with a goniometer. At the same time we found that the polarisation degree of the light reflected from the surface is so small that it cannot be reliably determined. Hence, the input data for pose estimation are limited to intensity, edges, and depth.

For our evaluation, we attached the rubber with its lateral surface to the goniometer table and oriented it in two different poses relative to the camera. The angular difference between the two poses is only a few degrees (cf. Fig. 6). For the determination of the ground truth, we replaced the rubber for each pose by a chequerboard of known geometry. The chequerboard was attached to the goniometer table and its pose was estimated using the rig finder algorithm described in [13], which is based on a bundle adjustment approach for camera calibration purposes. Due to the simple cuboid shape of the rubber the chequerboard pattern could be aligned at high accuracy into the same direction as the lateral surfaces of the rubber, such that the chequerboard pose could be assumed to be identical with the pose of the rubber.

The results of this first experiment are shown in Table 2. The deviations for this rather simple object are only a few tenths of a degree for the rotation angles and a few tenths of a millimetre for the lateral translations. The translation in z is determined at an accuracy of about 4 mm (which is about an order of magnitude lower than the lateral accuracy) or 1 percent. This is a reasonable result, given that only monocular image data are available.

3.2. Oil cap

In the second experiment we regard an oil cap consisting of plastic material. Since due to its complex shape this object cannot be attached to the goniometer table in a reproducable manner, we determined the ground truth pose in this experiment based on a stereoscopic bundle adjustment tool which exploits manually established point correspondences between a rectified stereo image pair and the CAD model of the object. As in the first experiment, the goniometer was used to determine the intensity and polarisation angle reflectance functions R_I and R_Φ. The light reflected by the surface of the oil cap is partially polarised by 10–20 percent, such that the polarisation angle can be used in our pose estimation framework in addition to intensity, edges, and depth. The intensity images of the two regarded poses are shown in Fig. 7, illustrating that at some places especially near the right image border the edges are not well-defined, such that the pose estimation algorithm to a large extent has to rely on intensity and polarisation information. The comparison to the ground truth is shown in Table 3, demonstrating that the object pose can be determined at an accuracy of 1–2 degrees for the rotation angles, some tenths of a millimetre for the lateral translations, and several millimetres or about 1 percent for the object distance. We observed that small deviations of the rotation angles can be compensated by correspondingly adjusting the albedo factor, leading to a lower accuracy of the rotation angles, compared to the rubber example. Due to the somewhat ill-defined edges the pose estimation fails when only edge information is used, as no convergence of the minimisation routine is achieved.

Parameter	Pose 1	GT 1	Pose 2	GT 2
roll [ ]	13:3	13:5	16:7	16:3
pitch [ ]	18:2	18:9	18:6	19:7
yaw [ ]	59:4	58:6	59:2	58:5
t_x [mm]	3:6	3:2	2:8	2:5
t_y [mm]	2:3	2:3	1:3	1:7
t_z [mm]	451:5	454:3	457:5	453:9

Table 2. Estimated pose and ground truth (GT) for the rubber example.

Parameter	Pose 1	GT 1	Pose 2	GT 2
roll [ ]	233.2	234.5	230.7	232.1
pitch [ ]	1.3	2.3	0.9	2.4
yaw [ ]	57.3	55.2	56.8	56.0
t_x [mm]	14.7	14.7	15.0	14.8
t_y [mm]	2.1	2.8	2.0	2.5
t_z [mm]	512.9	509.2	512.7	509.2

Table 3. Estimated pose and ground truth (GT) for the oil cap example.

(a) Input image (pose 1) (b) Input image (pose 2)

Figure 8. Input intensity images for the door hinge example. Greylevels are displayed in logarithmic scale.

For the oil cap example, it is possible to directly compare our results to those of the monocular edge-based template matching method proposed in [24], since in that work the same object and the same CAD model are regarded. The deviation of the rotation angles estimated in [24] from the corresponding ground truth is typically around 1-2 degrees but may also become larger than 3 degrees. In contrast to the method described in this study, it is assumed in [24] that the distance to the object is known, i. e. only five rather than six degrees of freedom are estimated in [24]. On the other hand, that method does not require a-priori information about the object pose.

3.3. Hinge

In our third experiment we regard another automotive part, a door hinge, consisting of cast metal with a rough and strongly specular surface (cf. Fig. 8). For the pose we have chosen for our experiment, the light from the point light source is reflected directly into the camera. The Canny edge detector yields a very large number of edges (cf. Fig. 2), thus providing no reliable information about the object pose. As a consequence, our approach fails when we attempt to perform a pose estimation of the hinge based on the extracted edge information. Just like the rubber in our first experiment, the surface of the hinge does not perceivably polarise the reflected light. Hence, we only use intensity and depth data as input information for our algorithm. The obtained results illustrate that our algorithm also works in the absence of some of the input cues and that it is suitable for pose estimation of objects with a strongly specular surface.

In this experiment, the chequerboard method could not be used for determining the ground truth since the hinge could not be attached to the goniometer in a reproducable manner, such that it was not possible to place it in a known position relative to the chequerboard and the goniometer. Similarly, the bundle adjustment tool based on manually established point correspondences could not be used since unlike the oil cap, the hinge does not display well-defined corner points. Hence, we compare the estimated poses to the difference imposed by the two chosen goniometer settings, values which are given at high accuracy. The estimated pose differences and the corresponding ground truth values are shown in Table 4. Although not all four geometric, photometric, and depth cues are available, the obtained results are comparable to or better than those obtained in the previous experiments (some tenths of a degree for the rotation angles, some tenths of a millimetre for the lateral translation, and some millimetres for the object distance). Hence, our method behaves in a robust manner with respect to a strongly specular object surface and cluttered edge information.

Parameter difference	Result	GT
roll [ ]	4.15	4.23
pitch [ ]	2.06	1.69
yaw [ ]	0.22	0.58
t_x [mm]	0.71	0.06
t_y [mm]	1.88	2.33
t_z [mm]	3.82	0.16

Table 4. Estimated pose differences and ground truth for the hinge example.

4. Summary and conclusion

In this study we have presented a monocular pose estimation framework which is based on photometric, polarimetric, edge, and defocus cues. A correspondingly defined error function is minimised by comparing the observed data to their rendered counterparts, where an accurate rendering of intensity and polarisation images is performed based on the material-specific reflectance functions determined with a goniometer. If a certain cue cannot be reliably measured or does not yield useful information, it can be neglected in the optimisation procedure.

The experimental evaluation, performed at an effective pixel resolution of 0.4 mm, has shown an accuracy of our method of several tenths of a degree to 1 degree for the rotation angles, 1 mm or better for the lateral object translation, and several millimetres, corresponding to about 1 percent, for the distance to the object. This accuracy is comparable to or higher than that of the monocular template matching approach in [24] exclusively relying on edge information. This result is achieved despite the fact that our method additionally provides an estimate of the distance to the object while the method in [24] assumes that the object distance is known. At this point it is interesting to compare the accuracy of our monocular approach with that achieved by a multiocular method. As an example, for the stereo-based approach described in [25] a rotational accuracy of 1.5 degrees and a translational accuracy of 2.4 mm are reported for an industrial part located at a distance of 600–800 mm[1].

The depth from defocus method has turned out to be a useful instrument for the estimation of object depth in the close range at an accuracy of about 1 percent. We have demonstrated the usefulness of our method under conditions typically encountered in industrial quality inspection scenarios such as the assembly of complex parts, where the desired pose of the whole workpiece or part of it is given by the CAD data and the inspection system has to detect small differences between the actual and the desired pose.

Beyond depth from defocus, our pose estimation framework is open for depth data obtained e. g. by active range measurement. Hence, future work will involve the inclusion of such independently obtained depth data into the described system.

References

[1] F. J. Canny. A Computational Approach to Edge Detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986.

[2] S. Chaudhuri and A. N. Rajagopalan. Depth from Defocus. A Real Aperture Imaging Approach. Springer, 1999.

[3] P. d'Angelo and C. Wöhler. 3d reconstruction of metallic surfaces by photopolarimetric analysis. In Proc. Scandinavian Conference on Image Analysis, pages 689–698, 2005.

[4] P. d'Angelo and C. Wöhler. 3d surface reconstruction by combination of photopolarimetry and depth from defocus. In Pattern Recognition, Proc. 27th DAGM Symposium, LNCS 3663, pages 176–183, 2005.

[5] P. d'Angelo and C. Wöhler. Image-based 3d surface reconstruction by combination of sparse depth data with shape from shading and polarisation. In Symposium of ISPRS Commission III, Photogrammetric Computer Vision (PCV06), Bonn, Germany, 2006. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, volume XXXVI(3), pages 124–129, 2005.

[6] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In SIGGRAPH 1997, volume 3, pages 369–378, 1997.

[7] P. Decaudin. Cartoon looking rendering of 3D scenes. Research Report 2919, INRIA, 1996.

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM, 24(6):381–395, 1981.

[9] D. Gavrila and V. Philomin. Real-time object detection for “smart“ vehicles. In Proc. IEEE Int. Conf. on Computer Vision, pages 87–93, 1999.

[10] R. Haralick, H. Joo, C. Lee, X. Zhuang, V. Vaidya, and M. Kim. Pose estimation from corresponding point data. SMC, 19(6):1426–1446, 1989.

[11] B. K. P. Horn and M. J. Brooks. Shape from shading. MIT Press, Cambridge, MA, USA, 1989.

[12] K. Kraus, J. Jansa, and H. Kager. Photogrammetry, Vol. 2, Advanced Methods and Applications. Dümmler, Bonn, 1997.

[13] L. Krüger, C. Wöhler, A. Würz-Wessel, and F. Stein. In-factory calibration of multiocular camera systems. In W. Osten and M. Takeda, editors, SPIE Photonics Europe (Optical Metrology in Production Engineering), pages 126–137, Sept. 2004.

[14] A. Kuhl, C. Wöhler, P. d'Angelo, L. Krüger, and H.-M. Groß. Monocular 3D Scene Reconstruction at Absolute Scales by Combination of Geometric and Real-aperture Methods. In Pattern Recognition, Proc. 28th DAGM Symposium, LNCS 4174, pages 607–616, 2006.

[15] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. Artificial Intelligence, 31(3):355–395, 1987.

[16] D. G. Lowe. Fitting parameterized three-dimensional models to images. IEEE Trans. Pattern Analysis and Machine Intelligence, 13(5):441–450, 1991.

[17] S. K. Nayar and R. M. Bolle. Reflectance based object recognition. International Journal of Computer Vision, 17(3):219–240, 1996.

[18] H. Niemann and J. Hornegger. A novel probabilistic model for object recognition and pose estimation. International Journal of Pattern Recognition and Artificial Intelligence, 15(2):241–253, 2001.

[19] Y. Nomura, D. Zhang, Y. Sakaida, and S. Fujii. 3-d object pose estimation based on iterative image matching: Shading and edge data fusion. In Proc. Int. Conf. on Computer Vision and Pattern Recognition, pages 866–871, 1996.

[20] T. Q. Phong, R. Horaud, A. Yassine, and P. D. Tao. Object pose from 2-D to 3-D point and line correspondences. International Journal of Computer Vision, 15(3):225–243, 1996.

[21] B. Rosenhahn, T. Brox, D. Cremers, and H.-P. Seidel. A comparison of shape matching methods for contour based pose estimation. In Combinatorial Image Analysis, LNCS 4040, pages 263–276, 2006.

[22] B. Rosenhahn, C. Perwass, and G. Sommer. Pose estimation of free-form surface models. In Pattern Recognition, Proc. 25th DAGM Symposium, LNCS 2781, pages 574–581, 2003.

[23] W. Szczepanski. Die Lösungsvorschläge für den räumlichen Rückwärtseinschnitt. In Deutsche Geodätische Kommission, Reihe C: Dissertationen, Heft Nr. 29, pages 1–144, 1958.

[24] C. von Bank, D. Gavrila, and C. Wöhler. A visual quality inspection system based on a hierarchical 3d pose estimation algorithm. In Pattern Recognition, Proc. 25th DAGM Symposium, LNCS 2781, pages 179–186, 2003.

[25] Y. Yoon, G. N. DeSouza, and A. C. Kak. Real-time tracking and pose estimation for industrial objects using geometric features. In Proc. International Conference on Robotics and Automation, pages 3473–3478, 2003.

[1] We inferred a resolution of about 0.9 mm per pixel from the example images shown in [25].