Registering Depth Maps from Multiple Views Recorded by Color Image Sequences

Rüdiger Beß
Lehrstuhl für Mustererkennung (Informatik 5)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstr. 3
D-91058 Erlangen
telephone: +49 9131 85-7891
fax: +49 9131 303811
email: bess@informatik.uni-erlangen.de

Abstract

In this paper we outline a system to search and classify objects in an unknown environment. As a first step towards this goal a dense depth map of the complete object surface is computed by use of a stereo approach adapted to monocular color image sequences. The resulting depth data of the stereo algorithm are input of the judgement, registration and fusion step described in this paper.

After judging the depth values and removing the poor ones we use common lines of sight to identify depth values in consecutive depth images which refer to the same 3D–point. This allows to compute the registering transformation directly, no matching step of the 3D–data is necessary. Advantages of this new approach are the independence of the registration step from accuracy in calibration and the possibility to register depth images containing only depth values of a single plane.

In the fusion step a 3D accumulator is used to combine depth values belonging to the same 3D-point. In this step we integrate filtering of neighboring depth values to restrict surface resolution and data size to the maximal resolution of the resulting complete 3D map of the object surface. This enables an extremely efficient evaluation of the depth images.

The algorithms described in this paper allow to overcome relatively inaccurate calibration and to compute 3D data from monocular image sequences by passive stereo approaches without placing a calibration pattern in the scene.

Keywords: stereo vision, monocular image sequence, dense depth map, multiple views, 3D reconstruction

1 Introduction

One of the challenging goals in computer vision is to search and classify objects actively in an unknown environment. A wide range of application would be opened by achieving this goal, including but not restricted to assistance of handicapped people, household robots and repairing of machines.

Many real world objects are clearly determined solely by their geometric properties. For this reason the three-dimensional form is an important source of characteristic information on an object and enables a reliable classification of a wide range of object types.

The goal of the system outlined in this paper is to set up a system to compute a complete dense depth map of the surface of an object without changing the environment and thus acquiring full information about the form of an object for use in classification and searching.

Traditionally there are two classes of approaches for depth computation: active and passive [Nie90]. While *active approaches* influence the environment in which they work, for instance by projecting patterned light onto the scene or by use of laser beams in a laser range finder, *passive approaches* place only a sensor in the scene to acquire visual information. Examples for passive approaches are stereo algorithms and depth from X approaches, where X can be replaced by motion, shading, texture, focus or defocus.

It is true that in different types of *invasive approaches* a passive sensor is used, but the environment is changed by placing the object into a special setup. For example in [SF95] the object is placed on a turn table to achieve the conditions in accuracy required and in many photogrammetric approaches a calibration rig is build up around the object to determine the transformation between all images of an image sequence [LGUM94].

Active approaches and approaches where the environment of an object has to be changed are in the following referred as *invasive approaches* to enhance the difference to *active vision approaches*, where for instance the camera is moved to get new views but the scene itself is not changed. Since the system is to be used in a natural environment without assistance of a human operator the type of the approach is restricted from the outset to a non invasive one.

In the class of non invasive approaches some are restricted either to special surface types or can determine depth only while no depth discontinuities occur, e.g. shape from texture, shape from shading or shape from focus and shape from defocus. In contrast with these restrictions stereo approaches can be used on a wide range of object forms and surface types [DA89] [Fua93]. With known motion stereo approaches are in principle applicable directly to monocular image sequences. The advantage of a stereo approach over shape from motion algorithms is the much lower number of images necessary to compute the complete surface of an object. Since matching in case of unknown camera position requires a very low displacement between consecutive images shape from motion algorithms typically need more than 500 images compared to 40 for a stereo approach.

This leads to the conclusion that a stereo approach on monocular image sequences is best suited to enable depth computation in the described environment.

The main problem preventing the use of stereo in monocular image sequences up to now was the lack of a calibration problem causing inaccurate depth measurements. So the question addressed in this article is: How should we judge and combine inaccurate depth information obtained from different image pairs? This problem is split into three parts: Measuring the quality of depth information to skip erroneous values (judgement), computing the transformation between depth images each obtained from a pair of images (registration) and combining depth images when the transformation is known (fusion).

According to the task the paper is organized as follows: After the introduction the experimental setup and system structure is outlined. The main part is divided in three sections: computation and judgement of depth, registering depth images and fusing depth values. The first of these sections (Section 3) deals with the calculation of depth values and the quality measure derived from the approach used. The second (Section 4) describes the measurement of the transformation between depth images resulting from different stereo image pairs. The third (Section 5) explains the fusion of depth values belonging to the same world point. Finally, experiments and results are presented.

2 Experimental Setup and System Design

The setup for searching in an unknown environment has to comply with several preconditions. To search actively for an object requires to alter the view and thus to change the camera position. In our experimental environment a monocular camera is mounted to the hand of a robot. So arbitrary movements of the camera within the working space of the robot are possible. This allows to take images of an object from different viewing angles. A fixed stereo camera is supervising the scene to detect interesting objects and thus enables the determination of hypotheses about the location of an object specifying the starting position for the monocular camera. Figure 1 is showing this setup.

The system structure is designed to interact actively with the environment based on sensor data. For this purpose three feedback loops are coupled. The innermost feedback loop deals with data driven

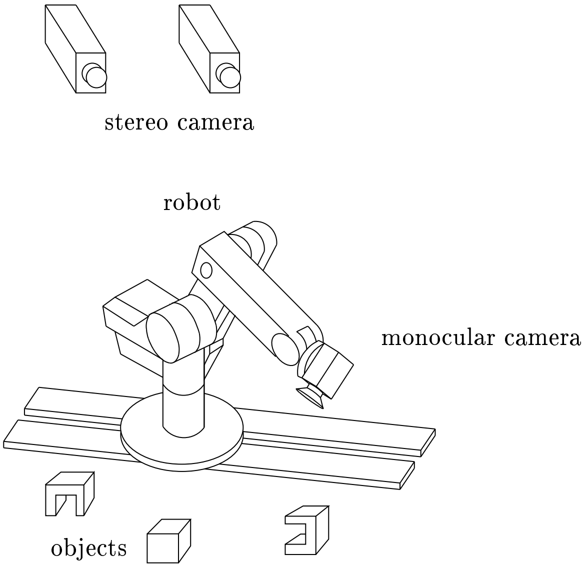


Figure 1: Setup showing the stereo camera supervising the scene and the robot with the monocular camera exploring one object

computation of depth data of an object, the task we concentrate on in this paper. The middle gives a classification result about the type and orientation of the object under consideration based on the depth data. The outermost is making hypotheses about the position of the object wanted. It is broken when the correct object is found.

Figure 2 gives an overview on the depth computation loop. First the monocular camera is moved to an initial position relative to an interesting object. Within the loop (Figure 2) a collision free path to the next camera position is determined, the camera is moved to this position and the exact position is measured. A depth image of the viewed object surface is computed with each two images, the depth images from different views are judged and integrated into a common depth map. After that the loop is starting again. The loop is broken either when a predefined number of images is taken or when the accuracy and completeness of the depth map exceeds a predefined value. The calibration, path planning and stereo algorithms are described more detailed in [Beß94, BPN96]. The next section is giving a brief overview as detailed as necessary for the explanation of the judgement, registration and fusion steps.

3 Depth Computation and Judgement

As mentioned in the introduction a passive stereo approach is best suited for active vision tasks in an unknown environment. Since we have only one moving camera the direct use of stereo algorithms is not possible. To adapt stereo algorithms to this setup the camera has to be calibrated and a robust matching step has to be developed dealing with the inaccuracies of calibration and intensity changes due to time and angular differences between consecutive images. For each image in a sequence the camera position is measured by the robot [Beß94], where a gripper–camera–calibration from Lenz and Tsai [TL88] is used to determine the camera position in a global coordinate system. The known camera position enables a combined feature– and correlation–based stereo approach. Thus a dense depth image is computed from each two color images recorded by the monocular camera [BPN96]. Each of this depth images contains depth information on one view of the object.

Since the camera position is known for each image, it is possible to determine depth from all images directly in a global coordinate system, without registering or fusion. But inaccuracies in the measurements of the camera position lead to a translation error in the relative position of two depth images up to 25 mm, with an average of 6.47 mm. The approach described in the subsequent sections is independent from this error, allowing to compute the transformation between the depth images very fast and accurate.

Before registration and fusion depth values are judged and erroneous ones are eliminated since the error would be propagated into the resulting depth map. Four parameters influencing the reliability of depth values are taken into consideration:

1. The *distance range* relative to the camera.
2. The *depth difference* of a depth value to its neighbors.
3. The *inclination* of a surface patch towards the optical axis.
4. The *common area* between consecutive depth images.

The first three parameters are motivated by the stereo approach used for depth computation. Its result — a partial dense depth map — is the input

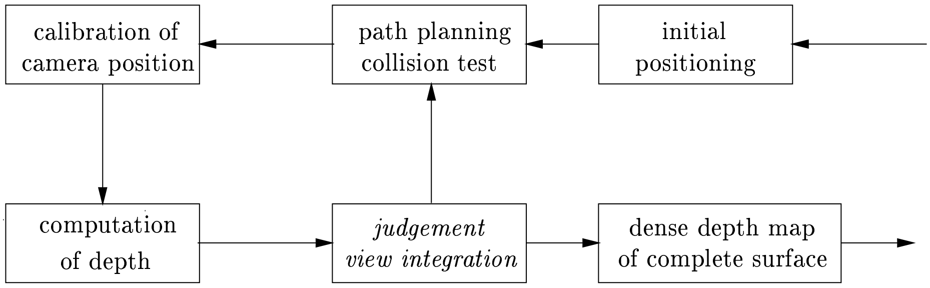

calibration of camera position
path planning collision test
initial positioning
computation of depth
judgement view integration
dense depth map of complete surface

Figure 2: Depth computation loop

for the combined feature and correlation based approach. The depth map is partial, because depth cannot be computed within homogeneous or occluded regions.

The camera must be as close as possible to the object to move the camera around it. Therefore the focus of the camera is adjusted to about 220 mm. The *distance range* relative to the camera is given by the depth of focus, the range where the object surface can be obtained without blurring. Outside this interval neither the edges necessary for the feature based matching step nor the correlation for the block matching step can be determined reliably.

The *depth difference* of a depth value to its neighbors allows the detection of outliers. Since depth computation is based on a block matching step, it is not possible to determine reliable depth values for surface patches which are smaller than a quarter of the block size. Therefore single depth values are judged by viewing the differences to the four neighboring values. If the difference against three of the four neighbors in a 4–neighborhood exceeds a threshold, the value is considered an outlier and rejected.

The restriction of the *inclination* of a surface patch towards the optical axis is necessary because a surface patch which is parallel or nearly parallel to the optical axis can not be measured reliably. One of the reasons is that such surfaces are often occluded in one of the two stereo images. Another reason is that the breadth of the edges decreases with decreasing inclination resulting in errors during feature detection and feature based matching. Therefore when the difference to one neighbor in an 8–neighborhood with distance n implies an inclination towards the optical axis above 60 degrees the value is rejected. The distance n must be chosen as a function of the depth resolution d_z and the resolution parallel to the x–axis of the depth image d_x. The result of $\arctan(d_z/(nd_x))$ must be below the threshold of 60 degrees, otherwise the least possible depth change would cause the rejection of the values at the border. Figure 3 illustrates this relation.

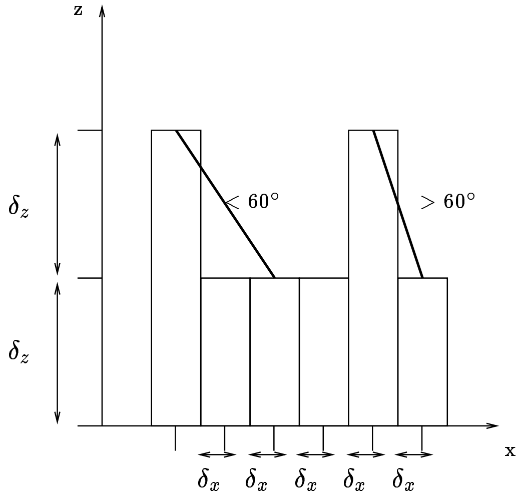


Figure 3: Minimal possible inclination at a depth change against a neighbor with distance two (enhanced line left) and distance one (enhanced line right)

The *common area* between consecutive depth images is measured after elimination of single erroneous depth values. A minimum size of this area is on the one hand necessary to ensure that the the number of common depth values is high enough to registrate the images. On the other hand this parameter allows to detect depth images with unusable quality. This typically occurs when the angular difference between two views is too high and the object is regular shaped and textured. If the object has several different surface patches with similar texture and shape the occlusion caused by the angular change leads to

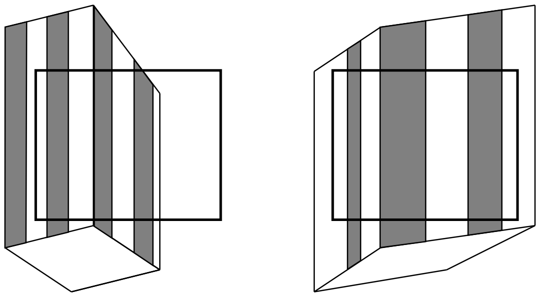


Figure 4: schematic example for a stereo image pair causing erroneous matches. The enhanced boxes show the visible part of the object in the right and left stereo image.

a completely wrong matching, leaving only surfaces with high inclinations or depth values outside the admissible distance range. A schematic example is shown in Figure 4. In both of the stereo images three stripes are visible but due to occlusion only one of them belongs to the same object surface patch.

4 Registering Depth Images

As mentioned in the introduction we have to overcome an error of up to 25 mm in the relative position of two depth images. Therefore it is insufficient to compose the depth data in a common coordinate system without further processing, although this would be possible by use of the known position of the stereo images relative to each other.

An usual approach to register overlapping depth images is to formulate it as an optimization problem: a transformation between each two depth images is searched which minimizes the error between overlapping regions [Kar93]. For this purpose a matching step has to be performed on the depth images identifying depth data which belong to the same 3D–point. This approach is not sufficient with the passive stereo approach used to determine the depth images. The noise is much higher and the resolution is lower than in active stereo approaches like those used there [Wah86, HR93]. Furthermore, if only one planar surface can be matched, the transformation between two depth images is not unique, the rotation around the normal of the planar surface and the translation parallel to it can not be determined. Due to the restrictions of the passive stereo approach this is likely to happen when used on simple regular workpieces like those build from rectangular parallelepipeds. In the worst case the image plane will be parallel to one side of the parallelepiped and since the angular difference between two consecutive views is less than 21°, the angular difference between the second view and any of the other sides will be more than 60°.

In general: Be δ the maximal inclination towards the image plane where a surface patch can be determined reliable (here 60°) and be α the angle between the optical axes in the two views of a stereo image pair. If $\delta + \alpha < \epsilon$ where epsilon is the minimal angle between one surface patch and any adjacent patch, then in the worst case depth can only be determined for this surface patch.

To overcome these problems we identify points in each two stereo images which belong to the same real 3D points instead of matching the 3D data. If we know that two image points in two different stereo images are projections of the same 3D point we also know that the two 3D positions computed from the image points must be equal. So we have to search for a transformation establishing the real relation between the depth maps.

We use the constraint that every image is part of at least two stereo images. The conversion of an image to a normalized stereo image is defined by a partial, bijective mapping of the coordinates from the image to the stereo image, partial, because the image section in general is slightly changing causing undefined regions in the normalized stereo image (see Figure 5). An image which is part of two stereo images thus defines a partial mapping between the coordinates of both stereo images. Since every image coordinate corresponds to a line of sight and every line of sight corresponds to one particular 3D–point for opaque objects — and only such are viewed — a mapping is established between equal depth values of two depth images.

This is illustrated in Figure 6, where ${}^{k}\boldsymbol{f}$ denotes the k-th image in an image sequence and ${}^{k,l}\boldsymbol{f}$ a stereo image pair computed from the images ${}^{k}\boldsymbol{f}$ and ${}^{l}\boldsymbol{f}$.

By use of this relation which is independent of the calibration accuracy, a system of equations can be formulated to compute the transformation between the depth images. In the following the equations defining this transformation are given.

4.1 Notation

As noted above ${}^{k}\boldsymbol{f}$ stands for the k-th image in an image sequence. Consequently upper left indices indicate the coordinate system of a vector: ${}^{k}\boldsymbol{p}$ denotes a point in the coordinate system of image ${}^{k}\boldsymbol{f}$. Lower

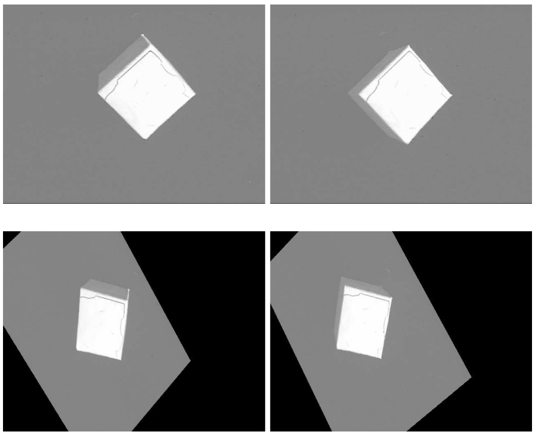


Figure 5: Two raw images (top) and corresponding normalized stereo image (bottom). Black areas are undefined.

left indices are used as labels for special points as ${}_k\boldsymbol{o}$, which signifies the origin of the coordinate system of ${}^k\boldsymbol{f}$. So ${}^l_k\boldsymbol{o} = (x, y, z)^T$ is the origin of image ${}^k\boldsymbol{f}$ in coordinate system of ${}^l\boldsymbol{f}$. Upper right and left indices are used as usual either as mathematical operation indices ($\boldsymbol{p}^T$, $\boldsymbol{R}^{-1}$) or as positional information in a vector or sequence (p_x, $\boldsymbol{p}_i$).

A vector in the normalized image coordinate system $\widehat{l,(m)}$ is denoted by ${}^{\widehat{l,(m)}}\boldsymbol{p} = ({}^{\widehat{l,(m)}}p_x, {}^{\widehat{l,(m)}}p_y, {}^{\widehat{l,(m)}}p_z)$. Where $l, (m)$ indicates that this coordinate system refers to a stereo image pair computed from the images ${}^l\boldsymbol{f}$ and ${}^m\boldsymbol{f}$, the brackets around m indicate that $l, (m)$ refers to the coordinate system of the image computed from ${}^l\boldsymbol{f}$ and finally the hat above ($\widehat{\ }$) indicates that the stereo image is normalized.

4.2 Transformation between raw image and normalized stereo image

In a normalized stereo image pair the axes of the two image coordinate systems are collinear and directed equally, the x-axes are positioned on the same straight line. Therefore the epipolar line of one image point is intersecting the image planes of the two stereo images in the same image scan line. This simplifies the matching step and leads to a very simple relation between disparity and depth. The relation between the coordinate systems of the raw images and the normalized stereo image pair is shown in Figure 7, where $\widehat{L}$ is a short form of $\widehat{l,(m)}$ and $\widehat{M}$ a short form of $\widehat{(l),m}$ denoting the normalized images

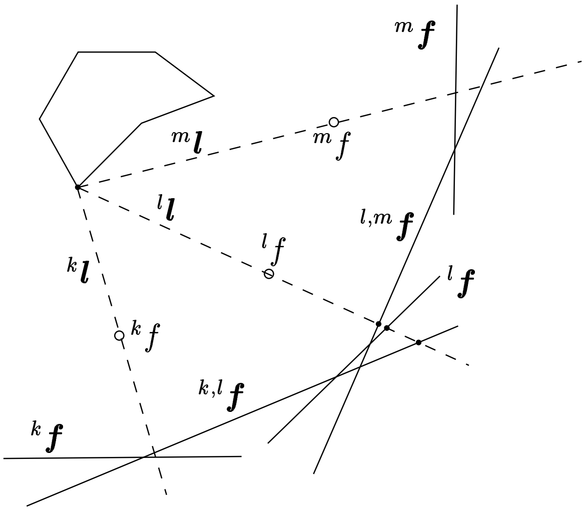


Figure 6: Common lines of sight. ${}^l\boldsymbol{l}$ is intersecting ${}^{k,l}\boldsymbol{f}$, ${}^l\boldsymbol{f}$ and ${}^{l,m}\boldsymbol{f}$.

computed from ${}^l\boldsymbol{f}$ and ${}^m\boldsymbol{f}$ respectively belonging to the stereo image pair ${}^{l,m}\boldsymbol{f}$.

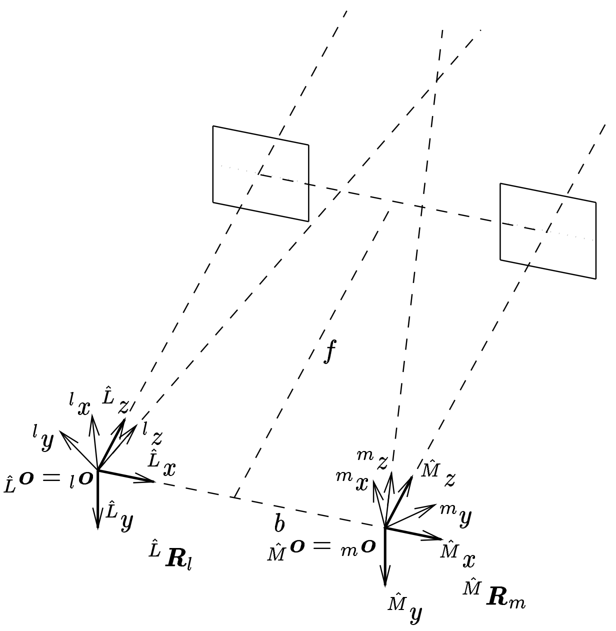
The parameters necessary to compute the relation between the normalized stereo image pair and the raw images are computed by the following steps:

- The internal camera parameters are determined by camera calibration [Pos90, Tsa86, Len87, Beß94]. Internal parameters are the focal length f, the distortion coefficients κ_1, κ_2, the pixel size δ_x, δ_y and the principal point $(C_x, C_y)^T$. These parameters do not change within an image sequence.

- The external camera parameters are determined from the robot position measurements by a gripper – camera – calibration [TL88, Beß94]. The external parameters define the camera position in the world coordinate system, given by the position of the optical centers ${}^w_l\boldsymbol{o}$, ${}^w_m\boldsymbol{o}$ and the rotation matrices ${}^w\boldsymbol{R}_l$ and ${}^w\boldsymbol{R}_m$ of the images ${}^l\boldsymbol{f}$ and ${}^m\boldsymbol{f}$ in the world coordinate system w.

- The transformation matrices into normalized stereo images ${}^{\widehat{L}}R_l$ and ${}^{\widehat{M}}R_m$ are computed from the camera position of the images ${}^l\boldsymbol{f}$ and ${}^m\boldsymbol{f}$ according to [Pos90].

With these parameters the following equations define the transformation between the normalized stereo image pair ${}^{\widehat{l,m}}\boldsymbol{f}$ consisting of ${}^{\widehat{l,(m)}}\boldsymbol{f}$ and ${}^{\widehat{(l),m}}\boldsymbol{f}$ and the raw images ${}^l\boldsymbol{f}$ and ${}^m\boldsymbol{f}$:



^{l}x
$^{\hat{L}}z$
^{l}y
^{l}z
$^{\hat{L}}x$
$_{\hat{L}}\boldsymbol{o} = {}_{l}\boldsymbol{o}$
$^{\hat{L}}y$
$^{\hat{L}}\boldsymbol{R}_l$
f
b
^{m}z
^{m}x
$^{\hat{M}}z$
^{m}y
$_{\hat{M}}\boldsymbol{o} = {}_{m}\boldsymbol{o}$
$^{\hat{M}}x$
$^{\hat{M}}y$
$^{\hat{M}}\boldsymbol{R}_m$

Figure 7: Relation between raw images and normalized stereo image pair

First pinhole coordinates $({}^{\widehat{L}}_{p}X, {}^{\widehat{L}}_{p}Y)$ are computed from the computer coordinates $({}^{\widehat{L}}_{c}X, {}^{\widehat{L}}_{c}Y)$. (Upper-case letters like $(X,\ Y)$ denote 2D–coordinates in the image plane, while lower–case letters $(x,\ y,\ z)$ are standing for 3D–coordinates).

$$\begin{pmatrix} {}^{\widehat{L}}_{p}X \\ {}^{\widehat{L}}_{p}Y \end{pmatrix} = \begin{pmatrix} ({}^{\widehat{L}}_{c}X - C_x)\delta_x \\ ({}^{\widehat{L}}_{c}Y - C_y)\delta_y \end{pmatrix} \tag{1}$$

The pinhole coordinates are mapped to 3D–coordinates in the image coordinate system of ${}^{l}\boldsymbol{f}$.

$$\begin{pmatrix} {}^{l}x \\ {}^{l}y \\ {}^{l}z \end{pmatrix} = {}^{\widehat{L}}R_l^{-1} \begin{pmatrix} {}^{\widehat{L}}_{p}X \\ {}^{\widehat{L}}_{p}Y \\ f \end{pmatrix} \tag{2}$$

The 3D–coordinates are projected into the image plane of ${}^{l}\boldsymbol{f}$ resulting in pinhole coordinates.

$$\begin{pmatrix} {}^{l}_{p}X \\ {}^{l}_{p}Y \end{pmatrix} = \frac{{}_{l}f}{{}^{l}z} \begin{pmatrix} {}^{l}x \\ {}^{l}y \end{pmatrix} \tag{3}$$

The pinhole coordinates are mapped to distorted computer coordinates which are equal to the raw image coordinates of the image taken by the camera.

$$\begin{pmatrix} {}^{l}_{c}X \\ {}^{l}_{c}Y \end{pmatrix} = \begin{pmatrix} \frac{{}^{l}_{p}X(1+\kappa_1 r^2 \kappa_2 r^4)}{\delta_x} + {}_{l}C_x) \\ \frac{{}^{l}_{p}Y(1+\kappa_1 r^2 \kappa_2 r^4)}{\delta_y} + {}_{l}C_x) \end{pmatrix} \tag{4}$$

For shortness equations 1 to 4 are united to one transformation.

$$\begin{aligned} \begin{pmatrix} {}^{l}_{c}X \\ {}^{l}_{c}Y \end{pmatrix} &= {}^{l}T_{\widehat{L}} \left\{ \begin{pmatrix} {}^{\widehat{L}}_{c}X \\ {}^{\widehat{L}}_{c}Y \end{pmatrix} \right\} \\ &= {}^{l}T_{\widehat{l,(m)}} \left\{ \begin{pmatrix} {}^{\widehat{l,(m)}}_{c}X \\ {}^{\widehat{l,(m)}}_{c}Y \end{pmatrix} \right\} \end{aligned} \tag{5}$$

The transformation between $({}^{l}_{c}X, {}^{l}_{c}Y)$ and another stereo image pair depending on this image can be determined analogously leading to a transformation ${}^{l}T_{\widehat{(k),l}}$.

$$\begin{pmatrix} {}^{l}_{c}X \\ {}^{l}_{c}Y \end{pmatrix} = {}^{l}T_{\widehat{(k),l}} \left\{ \begin{pmatrix} {}^{\widehat{(k),l}}_{c}X \\ {}^{\widehat{(k),l}}_{c}Y \end{pmatrix} \right\} \tag{6}$$

Due to the definition of the transformations each two points $({}^{\widehat{(k),l}}_{c}X, {}^{\widehat{(k),l}}_{c}Y)$ and $({}^{\widehat{l,(m)}}_{c}X, {}^{\widehat{l,(m)}}_{c}Y)$ are the same, if the equation

$$\begin{aligned} {}^{l}T_{\widehat{(k),l}} \left\{ \begin{pmatrix} {}^{\widehat{(k),l}}_{c}X \\ {}^{\widehat{(k),l}}_{c}Y \end{pmatrix} \right\} &= \begin{pmatrix} {}^{l}_{c}X_1 \\ {}^{l}_{c}Y_1 \end{pmatrix} \\ = \begin{pmatrix} {}^{l}_{c}X_2 \\ {}^{l}_{c}Y_2 \end{pmatrix} &= {}^{l}T_{\widehat{l,(m)}} \left\{ \begin{pmatrix} {}^{\widehat{l,(m)}}_{c}X \\ {}^{\widehat{l,(m)}}_{c}Y \end{pmatrix} \right\} \end{aligned} \tag{7}$$

is valid. So the 3D–position computed for these points must be equal.

Now let

$$\widehat{{}_{(k),l}}{}^{w}\boldsymbol{p}(X,Y) = \boldsymbol{p}_i \quad \text{and} \quad {}^{w}_{\widehat{l,(m)}}\boldsymbol{q}(X',Y') = \boldsymbol{q}_j \tag{8}$$

be homogeneous world coordinates computed from ${}^{\widehat{(k),l}}\boldsymbol{f}$ at coordinates (X,Y) and ${}^{\widehat{l,(m)}}\boldsymbol{f}$ at (X',Y').

Let $\boldsymbol{p}_i$, $\boldsymbol{q}_j$ with $i,j = 1,...,M$ be homogeneous coordinates and for $i = j$ let

$$\begin{pmatrix} {}^{\widehat{(k),l}}_{c}X \\ {}^{\widehat{(k),l}}_{c}Y \end{pmatrix} = {}^{l}T^{-1}_{\widehat{(k),l}}\, {}^{l}T_{\widehat{l,(m)}} \left\{ \begin{pmatrix} {}^{\widehat{l,(m)}}_{c}X' \\ {}^{\widehat{l,(m)}}_{c}Y' \end{pmatrix} \right\} \tag{9}$$

be valid.

Then a linear transformation

$$\boldsymbol{H} = {}^{\widehat{(k),l}}\boldsymbol{H}_{\widehat{l,(m)}}$$

can be determined by minimizing the mean square error

$$\sum_{i=1}^{M}(\boldsymbol{q}_i - \boldsymbol{H}\boldsymbol{p}_i)^2 = \epsilon$$

The depth values obtained from the normalized stereo images $^{\widehat{(k),l}}\boldsymbol{f}$ and $^{\widehat{l,(m)}}\boldsymbol{f}$ are then registrated by the following function:

$$\widehat{\boldsymbol{p}}_i = \begin{cases} (\boldsymbol{H}\boldsymbol{p}_i + \boldsymbol{q}_i)/2 & \text{if } \boldsymbol{p}_i \text{ and } \boldsymbol{q}_i \text{ defined} \\ \boldsymbol{H}\boldsymbol{p}_i & \text{if only } \boldsymbol{p}_i \text{ defined} \\ \boldsymbol{q}_i & \text{if only } \boldsymbol{q}_i \text{ defined} \end{cases} \tag{10}$$

The equations above define an exact mapping between image coordinates in a raw image and the normalized stereo images. In general integral coordinates in a raw image will not be mapped to integral coordinates in the two normalized stereo image computed from it, but we know depth only at integral coordinate values of the normalized stereo image. To overcome this drawback two approaches are possible: on the one hand we can interpolate depth values for the non integral coordinates, on the other hand we can round the coordinates to the next integral value. So what is the maximal error caused by the later?

Since the maximal inclination of a plane in the depth image is restricted to 60 degrees and the maximal distance between the real and the rounded value is half a pixel, the maximal error is $2d_x \tan(60°)$, where d_x is the minimal distance in x-direction between two depth values. In our environment the maximal error is 0.87 mm, which is lower than the depth resolution of about 1.00 mm. Neglecting this error allows to compute a function table during the interpolation step defining the mapping between integral coordinates in the raw image and rounded coordinates in the normalized stereo images. Thus the effort to registrate the depth values is reduced to two accesses to this function tables per depth value and the evaluation of the linear equation system.

The resulting linear transformation between two depth images is restricted to translation and rotation. If scaling is allowed a depth image with only one planar surface patch will lead to a transformation adapting the depth values to the error perpendicular to this planar surface.

By the registration algorithm described the mean distance between depth values in consecutive images is reduced from 6–24 mm to less than 2 mm.

Since only consecutive images are linked by this transformation errors in the translation can be cumulative in an image sequence. So far this did not cause problems in registration or fusion but a relaxation algorithm to optimize all transformations in an image sequence simultaneously will be the next step to exclude the possibility.

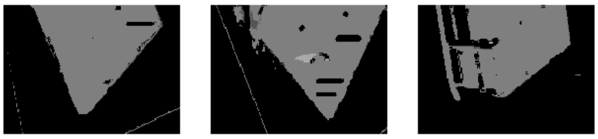


Figure 8: Three consecutive depth images of one corner of a cube, in the middle one a part is undefined.

5 Fusing Depth Values

For overlapping regions in consecutive depth images the fusion problem is solved by the algorithm outlined in the previous section, the relation between the depth values is established by the transformation given there. Depth values where this relation is unknown occur typically when the value for one 3D–point is undefined in the middle of three consecutive depth images, while it is defined in the first and third, an example is shown in Figure 8. Since the angular difference between the first and fourth image exceeds 60 degrees it is very unlikely that a value is undefined in two consecutive depth images but defined in the preceding and succeeding one.

We use the Euclidean distance between the points in the registered depth map to identify such values. If the distance between two points belonging to different depth images is below the average distance error ϵ — in our environment 2 mm —, and there is no other point in the same depth image with lesser distance the two points are fused. Formally:

$$\text{If } {}^{k}\boldsymbol{p}_i \text{ nearest neighbor to } {}^{l}\boldsymbol{p}_j \text{ and } \|{}^{k}\boldsymbol{p}_i - {}^{l}\boldsymbol{p}_j\| < \epsilon$$

$$\text{and } \neg\exists\ {}^{k}\boldsymbol{p}_a : \|{}^{k}\boldsymbol{p}_a - {}^{l}\boldsymbol{p}_j\| < \|{}^{k}\boldsymbol{p}_i - {}^{l}\boldsymbol{p}_j\|$$

$$\text{then } {}^{k}\tilde{\boldsymbol{p}}_i = {}^{l}\tilde{\boldsymbol{p}}_j = ({}^{k}\boldsymbol{p}_i + {}^{l}\boldsymbol{p}_j)/2$$

An example is shown in Figure 9. While ${}^{m+2}\boldsymbol{p}_{j+1}$ is the nearest point in the depth image ${}^{m+2}_{3D}\boldsymbol{f}$ with respect to ${}^{m}\boldsymbol{p}_{i+2}$, ${}^{m}\boldsymbol{p}_{i+1}$ is nearer to ${}^{m+2}\boldsymbol{p}_{j+1}$. So only ${}^{m+2}\boldsymbol{p}_{j+1}$ and ${}^{m}\boldsymbol{p}_{i+1}$ are fused.

In principle the complexity of this approach is quadratic, the number of comparisons necessary is equal to the product of the number of depth values in the two depth images. The distance constraint allows to reduce this effort by ordering the depth values and compare against values within the maximal

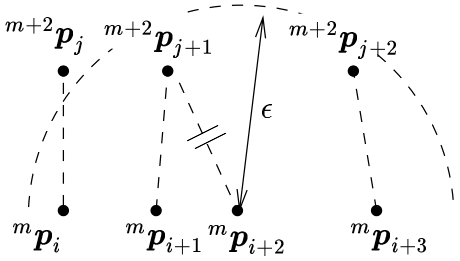

$^{m+2}\boldsymbol{p}_j$
$^{m+2}\boldsymbol{p}_{j+1}$
$^{m+2}\boldsymbol{p}_{j+2}$
ϵ
$^{m}\boldsymbol{p}_i$
$^{m}\boldsymbol{p}_{i+1}$
$^{m}\boldsymbol{p}_{i+2}$
$^{m}\boldsymbol{p}_{i+3}$

Figure 9: Example for the fusing algorithm

distance only. To make efficient use of this distance constraint the points have to be sorted according to their position. Here we use the fact that the volume is restricted to the working space of the robot and the depth of focus of the camera. Thus we can map this volume to a finite set of discrete entries in a three-dimensional array. The number of pixels in x-direction n_x in the normalized stereo image is one upper bound for the number of entries necessary in one dimension. A second boundary is given by the number of pixels in x-direction times the resolution in x-direction d_x divided by the depth resolution d_z. The depth resolution is approximately e/bd_x where e is the distance from focus to object and b is the stereo base, thus normally $d_z > 3d_x$, like in our environment with $d_x = 0.25$ mm and $d_z = 1.00$ mm. This is further reduced by the fact that the accuracy is lower than the minimal depth difference, the error remaining after registration is at least two times as high. Therefore a typical value for the maximal number of entries in one dimension is 64, the total number of entries $64^3 = 262144$ allowing a combination of registration, fusion and adaption of resolution in one step. The algorithm for this step is shown in Figure 10.

Results are given in the next section.

6 Experiments and Results

The algorithms were implemented using HIPPOS [Pau92, PH95], a NIHCL based object oriented class library designed for image analysis. For several objects the camera was positioned to record 40 views each. One result is shown in Figure 11. The pictures show 6 disparity images of a Rubic's cube (top and middle row). In this images dark grey-levels indicate far points and light grey-levels near points with respect to the camera. Black regions show undefined values. The 3 images in the bottom row show 3 views

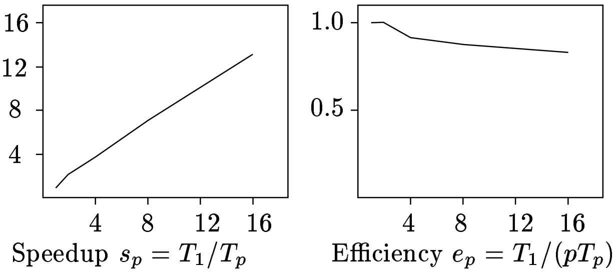


Figure 12: Speedup and efficiency of parallelization by data partitioning

of the resulting three-dimensional after registration before fusion and filtering.

Computationally expensive parts of the stereo algorithm are parallelized on MEMSY, a modular, expandable multiprocessor system. This system consists of 20 nodes with each 4 processors 88100 (24 mflops, 12 mflopds per processor). The speedup and efficiency is given in Figure 12.

For object recognition and modelling a segmentation step based on the computed 3D–data will be subject to further work.

7 Conclusion

The concept of common lines of sight allows a very fast and accurate registering of consecutive depth images obtained by a stereo algorithm adapted to color image sequences. No 3D–matching step is required to register the depth images thus allowing to register even single planar surface patches. Combining registering, fusion and data reduction leads to a very fast algorithm, able to compute the 3D image from 40 depth images in less than 10 seconds. The algorithms outlined in this paper enable the computation of depth with a relatively inaccurate calibration thus building a procedure to compute depth from monocular color image sequences successfully by a passive non invasive stereo approach.

8 Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft under grant number SFB 182. Only the author is responsible for the contents.

```
$l$: edge length of an entry in the 3D array
$n$: number of entries in one dimension
compute extremal values for $x$-, $y$-, $z$-dimension
determine $x_{min}$, $x_{max}$, $y_{min}$, $y_{max}$, $z_{min}$, $z_{max}$ from all registered 3D point values $\boldsymbol{p}$
compute the center of the object
$x_{mid} = (x_{max} - x_{min})/2 \quad y_{mid} = (y_{max} - y_{min})/2 \quad z_{mid} = (z_{max} - z_{min})/2$
FOR all registered 3D point values $\boldsymbol{p} = (x, y, z)$
    compute the appropriate index in the 3D object array
    $i_x = (x - x_{mid})/l + n/2 \quad i_y = (y - y_{mid})/l + n/2 \quad i_z = (z - z_{mid})/l + n/2$
    update the center of gravity and the number of points in this entry
    $p_{i,k,m} = (p_{i,k,m} * n_{i,k,m} + p)/(n_{i,k,m} + 1)$
    $n_{i,k,m} = n_{i,k,m} + 1$
FOR all entries in the 3D–array
    IF current entry is not empty
    THEN FOR all neighbors of this entry
            IF current neighbor is not empty
            THEN compute the distance between the center of gravity of this neighbor and the actual entry
                 IF computed distance is below threshold $\theta$ with $\theta < l/2$
                 THEN fuse both entries by computing the common center of gravity
```

Figure 10: Fusion of 3D point values in a 3D array

References

[Beß94] R. Beß. Kalibrierung einer beweglichen, monokularen Kamera zur Tiefengewinnung aus Bildfolgen. In W. G. Kropatsch and H. Bischof, editors, *Tagungsband Mustererkennung 1994*, volume 5 of *Informatik Xpress*, pages 524 – 531, Berlin, 1994. Springer.

[BPN96] R. Beß, D. Paulus, and H. Niemann. 3d recovery using calibrated active camera. In *International Conference on Image Processing*, Lausanne, Sept. 1996. IEEE.

[DA89] U.R. Dhond and J.K. Aggarwal. Structure from stereo: A review. *IEEE trans. on Systems, Man and Cybernetics*, 19(6):1489–1510, 1989.

[Fua93] P. Fua. Combininig stereo and monocular information to compute dense depth maps that preserve depth discontinuities. *Machine Vision and Applications*, 6:35–49, 1993.

[HR93] G. Häusler and D. Ritter. Parallel three-dimensional sensing by color-coded triangulation. *Applied Optics*, 32(35):7164–7169, Dec. 1993.

[Kar93] S. Karbacher. Surface reconstruction from multiple range images. In Lehrstuhl für Optik, editor, *Annual Report*, page 119, Erlangen, 1993.

[Len87] R. K. Lenz. Linsenfehlerkorrigierte Eichung von Halbleiterkameras mit Standardobjektiven für hochgenaue 3D-Messungen in Echtzeit. In E. Paulus, editor, *Proceedings 9. DAGM-Symposium*, Informatik Fachberichte 149, pages 212–216, Berlin, 1987. Springer.

[LGUM94] F. Leberl, M. Gruber, P. Uray, and F. Madritsch. Trade–offs in the reconstruction and rendering of 3–D objects. In W. G. Kropatsch and H. Bischof, editors, *Tagungsband Mustererkennung 1994*, volume 5 of *Informatik Xpress*, pages 58 – 73, Berlin, 1994. Springer.

[Nie90] H. Niemann. *Pattern Analysis and Understanding.* Springer, Berlin, 1990.

[Pau92] D.W.R. Paulus. *Objektorientierte und wissensbasierte Bildverarbeitung.* Vieweg, Braunschweig, 1992.

[PH95] D. Paulus and J. Hornegger. *Pattern Recognition and Image Processing in C++.* Advanced Studies in Computer Science. Vieweg, Braunschweig, 1995. second edition to appear 1997.

[Pos90] Stefan Posch. *Automatische Tiefenbestimmung aus Grauwert-Stereobildern.* Dissertation, Deutscher Universitäts Verlag, Wiesbaden, 1990.

[SF95] W. B. Seales and O. D. Faugeras. Building three–dimensional object models from image sequences. *Computer Vision and Image Understanding*, 61(3):308–324, May 1995.

[TL88] R. Y. Tsai and R. K. Lenz. Real time versatile robotics hand/eye calibration using 3d machine vision. In *International Conference on Robotics and Automation*, pages 554–561, Philadelphia, Pa., Apr. 1988. IEEE.

[Tsa86] R. Y. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. In *International Conference on Computer Vision and Pattern Recognition*, pages 364–374, Miami Beach, Florida, 1986. IEEE.

[Wah86] F. M. Wahl. A coded light approach for depth map acquisition. In G. Hartmann, editor, *Proceedings 8. DAGM-Symposium*, number 125 in Informatik Fachberichte, pages 12–17, Berlin, July 1986. Springer.

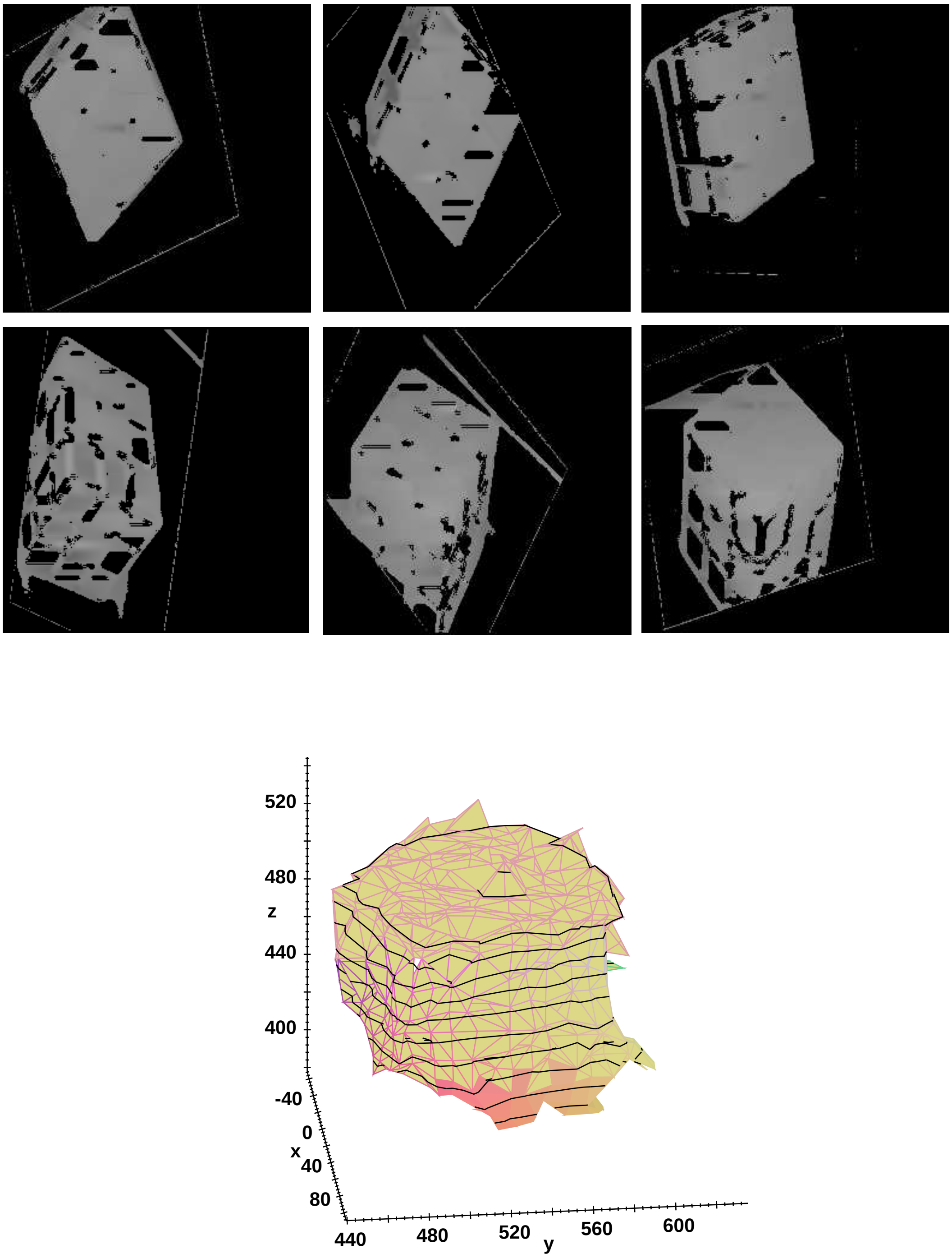


Figure 11: 6 depth images of a colored cube (top and middle row) and resulting depth data (bottom).