Image-Based Modeling and Its Application in Image Processing[1]

H. Niemann*, J. Denzler*, B. Heigl, F. Vogt*, S. Krüger***, W. Hohenberger***, and C. H. Schick*****

** Chair for Pattern Recognition (Informatik 5), University of Erlangen-Nüremberg, Martensstr. 3, 91058 Erlangen, Germany*

e-mail: {niemann,denzler,vogt}@informatik.uni-erlangen.de

*** Siemens Medical Solutions Inc., Forchheim, Germany*

**** Department of Surgery, University Hospital of Erlangen, Germany*

Abstract—This paper describes the recording and self-calibration of lightfields as an approach to purely image-based modeling of a scene or object. The application of lightfields is shown through the example of self-localization of a robot and as a support for laparoscopic surgery.

1. INTRODUCTION

There are different applications where a CAD-based modeling or some other type of symbolic modeling is difficult to achieve; an alternative is the use of a purely image-based type of model, where the model basically is a set of images. One image-based model is the lightfield, originally introduced in computer graphics to render photorealistic images of arbitrary scenes including, for example, flowers, fur, or hair. The lightfield is a four parameter representation of the plenoptic function.

A lightfield can be recorded by moving a handheld camera in front of, around, or inside of an object; in this case, no information about the camera pose is given and it must be recovered by self-calibration. It may also be recorded by having a robot moving the camera; if accurate position information about the robot hand is available, only the intrinsic camera parameters have to be determined, for example, in a preprocessing step by using a calibration pattern.

In this contribution, we describe a robust method for recording lightfields by a handheld camera. We point out the use of lightfields for self-localization of a moving robot and as a support for laparoscopic surgery.

2. LIGHTFIELDS

For any point in space, the plenoptic function gives the light intensity emitted in some direction. Hence, it is a five-parameter function. The lightfield is a four-parameter representation of light rays where in the (s, t) plane we have camera positions and in the (u, v) plane, pixel coordinates [4]. To overcome the limitation of having cameras (view points) only in a plane, the freeform lightfield allows camera positions at arbitrary locations. Basically, a freeform lightfield is a sequence of images taken at different camera positions plus the extrinsic parameters (translation, rotation) of the camera. By a suitable viewer (or renderer) a freeform lightfield may be used to generate a view from a position or viewpoint which is different from the recording positions. The advantage of the lightfield is that it gives a purely image-based model of an object or a scene. The challenges associated with lightfields are the accurate recording, the fast rendering, the coding suitable for rendering, and the efficient usage.

If a (freeform) lightfield is recorded by moving a handheld camera around an object, in a scene, or within, e.g., the stomach, the camera parameters have to be obtained from calibration. This is done by computing and tracking feature points in the images of the sequence, computing an initial factorization over a few images of the sequence, and extending the factorization to the whole sequence. The steps for calibration of an initial sequence are the following [3]:

- apply the weak-perspective factorization method;
- if applicable, eliminate outliers using LmedS;
- create a reconstruction of poses of perspective cameras with a roughly estimated focal length and with the image center as the principal point;
- perform a nonlinear optimization of this solution by alternately optimizing the parameters of the cameras and the coordinates of the scene points;
- use this reconstruction to determine projective depths and apply them to the perspective factorization method;
- perform self-calibration by an absolute quadric [6];
- if the recording camera can be assumed to have more or less constant intrinsic parameters, then

[1] This article was submitted by the authors in English.

Received December 1, 2003

Pattern Recognition and Image Analysis, Vol. 14, No. 2, 2004, pp. 184–189.

improve the self-calibration with the method described in [2];

• apply nonlinear optimization of scene points and camera parameters either by assuming constant intrinsic parameters or by estimating the camera parameters independently from each other.

In order to get results for a long sequence of images, e.g., taken from a circular view around an object, the initial sequence has to be extended.

Figure 1 shows a result of this process. A camera was moved around a head, one image and the detected scene points are shown on the top of the figure. The camera positions computed by the above-described approach and the set of tracked 3D scene points are shown at the bottom.

Fig. 1. One image out of an image sequence (top) and the computed camera positions (bottom).

3. SELF-LOCALIZATION

Self-localization is one of the most important tasks that a mobile system has to solve. Although odometry is quite an accurate source of information for small local movements of a robot, for larger motion trajectories, odometry fails, since small errors for each time step accumulate over time [9]. Thus, a robot must permanently sense the environment in which it is moving to correct its position estimate by matching its model of the world with the sensor data recorded. The main issue, besides the localization itself, is the choice of the model of the world and its automatic reconstruction. The problem of global self-localization, i.e., the localization in a scene without any *a priori* information also remains unsolved.

Despite the fact that visual sensors provide more information than classical robot sensors, state-of-the-art robot self-localization is based on sonar sensors and laser range finder [9, 10]. The model of the environment is either created manually by using a CAD model or by an extra step called map building. The combined approach to online map building and self-localization, called self-localization and mapping (SLAM) in the robotics literature, is one of the most difficult problems in robotics in general, although several dedicated solutions exist, that solve SLAM in certain applications [10].

With the lightfield a new kind of model for scenes is available, that can be applied to vision-based self-localization. Its benefits are as follows:

• The model can be automatically reconstructed from image sequences taken by a handheld camera (compare Sec. 2). Thus, even for environments where no CAD model is available or it is difficult or even impossible to reconstruct such a model automatically from sonar or laser data (or manually by a user), the model can be built in the case of a lightfields without user interaction.

• The lightfield allows the rendering of photorealistic images taken from an arbitrary viewpoint in the scene. As a consequence, for any position hypothesis, which the robot has computed, the corresponding sensor data (i.e., the image), that the robot should acquire, can be virtually created using the lightfield. As a consequence, the update of its position estimate is possible based a pair of images: an image that should be seen and an image that is actually recorded.

• Due to its image-based representation of the scene, a lightfield is a perfect means for photorealistic modeling. Photorealism of such a model is one important advantage that makes it suitable for vision-based self-localization. In contrast to feature-based self-localization, the information reduction step is avoided and all the information in the image is maintained, for example, reflections.

In our case, probabilistic robot self-localization is done based on the equations

$$\mathbf{q}_t^* = \operatorname{argmax}_{q_t} p(\mathbf{q}_t | H_t) H_t$$
$$= [\mathbf{m}_t, \mathbf{o}_t, \mathbf{m}_{t-1}, \mathbf{o}_{t-1}, \ldots, \mathbf{m}_o, \mathbf{o}_o]$$

with $p(\mathbf{q}_t|H_t) = p_{q,H}$ the probability density of the robot being in state $\mathbf{q}_t$ given a history H_t of actions $\mathbf{m}_t$ followed by observations $\mathbf{o}_t$. The desired robot position and heading are represented in its state $\mathbf{q}_t$. The density

$$p_{q,H} = \frac{1}{c} p(\mathbf{o}_t | \mathbf{q}_t) \int \varphi(\mathbf{q}_t, \mathbf{q}_{t-1}, H_t) d\mathbf{q}_{t-1}$$

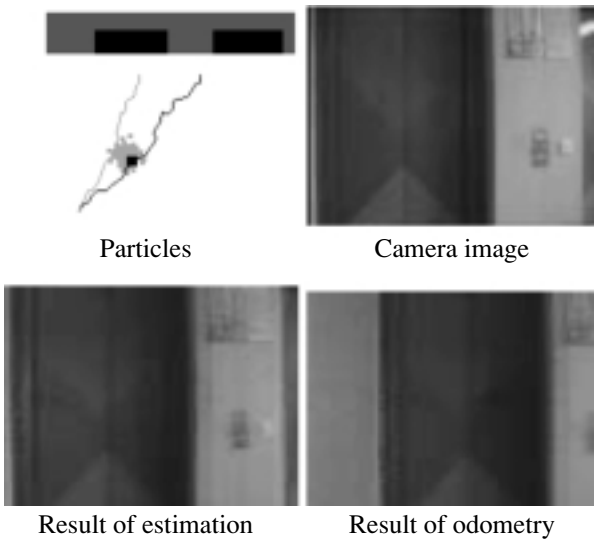


Fig. 2. Result of self-localization with a perturbed path.

Fig. 3. Setup during a minimal-invasive operation in a modern operation room. Two surgical instruments and the endoscope are introduced into the abdomen. Here, the endoscope is moved by a speech-controlled robot arm. A camera at the end of the endoscope provides the image of the operation area. Two additional monitors can be used to display improved images or 3D rendering.

is propagated over time by a particle filter [11] using

$$\varphi(.) = p(\mathbf{q}_t|\mathbf{q}_{t-1}, \mathbf{m}_t)p(\mathbf{q}_{t-1}|H_{t-1})$$

The crucial point of such a Bayesian approach is the so-called likelihood function $p(\mathbf{o}_t, \mathbf{q}_t)$, describing the relationship between the state and the observation at a certain time step t. This density explicitly models uncertainty in the sensing process. Having the lightfield as a model of the scene, this density can be easily defined as a comparison between the expected observation (image) based on the state estimate and the true observation made by the mobile system. One straightforward but nonetheless good way for defining $p(\mathbf{o}_t|\mathbf{q}_t)$ is

$$p(\mathbf{o}_t|\mathbf{q}_t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{(\mathbf{o}_t - \hat{\mathbf{o}}_t(\mathbf{q}_t))^2}{2\sigma^2}\right)$$

with $\hat{\mathbf{o}}_t(\mathbf{q}_t)$ being the synthetic image rendered from viewpoint given by the state estimate $\mathbf{q}_t$ and the operator "-" in the exponent being any distance function between two images, for example, a pixelwise difference operation. The variance σ^2 can either be estimated from examples or set empirically. A small distance between the two images is a strong hint for the estimated state being correct. With this likelihood function, the particle filter framework is completed and sequential estimation can be performed by using the local (and quite accurate) odometry information $\mathbf{m}_t$ for constructing the motion model $p(\mathbf{q}_t|\mathbf{q}_{t-1}, \mathbf{m}_t)$. Again, some assumptions must be made for the density, for example, a Gaussian with mean $(\mathbf{q}_{t-1} + \mathbf{m}_t)$ and a variance that must be estimated from examples.

Figure 2 shows a result for self-localization using a lightfield as a model and a particle filter for state estimation. The top left image shows the true motion path towards the elevator doors and the particle set for one time step, correctly following the true trajectory (dark line). In contrast to that, the odometry information would result in a wrong trajectory estimate (light line). The images (top right, bottom left and right) show for the time step, indicated in the top left image by the cloud of particles, the following:

- the image (top right) taken by the camera, which is used as observation for the particle filter,
- the rendered synthetic image based on the state estimate (bottom left), and
- the image (bottom right), also rendered from the lightfield, that would result from the pure odometry information.

This result clearly indicates that the estimated position by a particle filter using the lightfield is fairly accurate compared to the estimation based on odometry information only.

4. VR IN LAPAROSCOPIC SURGERY

Laparoscopic surgery is carried out by the surgeons without direct visual contact to the operation situs inside the abdomen. Instead, the video of the operation area is displayed on a monitor for visual feedback (see Fig. 3). Compared to the conventional operation with a large incision, the personal strain of the surgeon is increased: the image quality may be low due to image

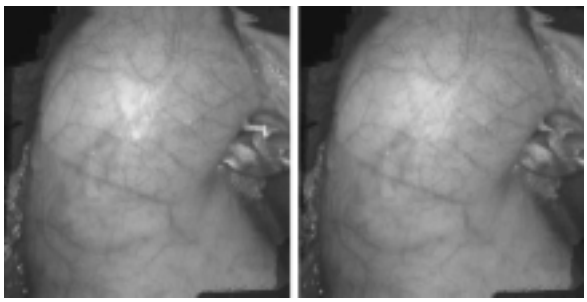
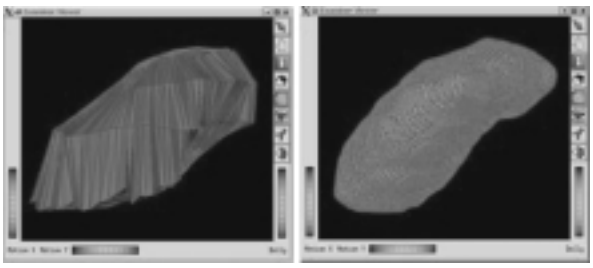


Fig.4. Rendered images of a lightfield of a gall (left) with specularity section replaced (right).



Gall, LF triangulation Gall, CT triangulation

Fig. 5. Triangular meshes from different modalities to prepare for registration.

degradations (lens distortion, smoke, small flying particles), the field of vision is limited, and almost no haptic feedback is available.

The goal in laparoscopic surgery is to support the surgeon by improving the image quality, by augmenting an image with other organs (e.g., obtained from preoperative CT images) or vessels (e.g., obtained from a database), and by providing a 3D recording and rendering of a surgery situation. By using lightfields as a scene model, a truly three-dimensional representation is available. The (new) setup in the operation room for implementing this goal is shown in Fig. 3.

In our experiments and evaluations [12], it has been considered significantly useful by several surgeons to process the image sequence recorded during surgery by a color median filter over time, employ color normalization, compute a geometric correction, and substitute specularities by using images rendered from the lightfield with a (virtual) camera position, where no or reduced specularities are present [7, 8]. Figure 4 shows an example of this substitution.

The conventional way of lightfield reconstruction (cf. Sec. 2) has its limits in laparoscopic surgery. Only if the following prerequisites are met, a reasonable result can be achieved:

• As little movement as possible in the recorded scene. It may be noted that (some) movement by heartbeat and respiration will always occur.

• As smooth camera motion as possible: since the operation situs is very close to the lens (5 to 25 cm), even small movements of the endoscope result in large movements in the camera image.

• As high image quality as possible: the correctness of feature tracking depends on the image quality.

• As "good" scene as possible: prominent points are selected as features that can be tracked. Homogeneous regions do not contain such points.

To overcome these limits, an endoscope positioning robot, e.g., AESOP (see Fig. 3), can be used for lightfield reconstruction. If the kinematics of the robot is known, the hand-eye transformation from the endoscope plug to the endoscope tip (real camera position) can be calculated by using a calibration pattern. Applying this method, the extrinsic camera parameters can be calculated for each image without limitations. As the intrinsic camera parameters do not change during the course of an operation, they can be determined before each operation by common camera calibration techniques.

The process of superpositioning the lightfield during surgery with a preoperative CT can be separated into three steps:

(1) Segmentation of CT data.

(2) Generation of triangular meshes.

(3) Registration of the triangular meshes.

Segmentation is done with a semi-automatic approach using seed points and a filling algorithm with a threshold. Triangular meshes are then generated from the depth information of the scene (contained in the lightfield) and the CT data, respectively (see Fig. 5).

The triangular meshes are the input for the registration process. The coarse registration is done interactively by selecting three corresponding 3D points in each data set. From these points the registration transformation can be calculated. A refinement of the registration is done by applying an Iterative-Closest-Point algorithm [13]. The visualization of the registered modalities is done by displaying two related windows. The first one contains the visualization of the lightfield, the second contains the visualization of the segmented CT data. The surgeon can navigate in each window and view the operation situs together with the segmented CT data from the corresponding direction.

CONCLUSIONS

Image-based modeling of objects and scenes requires the acquisition of calibrated image sequences which can be done by a handheld camera or a robot providing pose information. Robust self-calibration of a sequence from a handheld camera is possible with the approach outlined in this contribution. A useful representation is the freeform lightfield. The lightfield can be used to perform robot self-localization, to reduce specularities in laparoscopic images, and to augment such images by virtual objects (e.g., other organs or vessels)

obtained from preoperative CT images or from an anatomical database. Other applications, not mentioned in this contribution, are the use of lightfields for object tracking and for generating training views for object recognition.

To increase the versatility of the lightfield, dynamic lightfields are also needed for object changing with time due to motion or deformation. Work in this direction is in progress [5].

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), SFB 603, TP B6, and C2. Only the authors are responsible for the content.

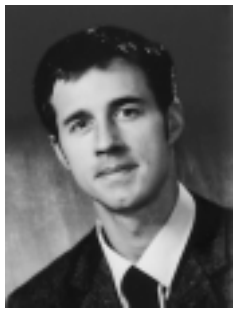
REFERENCES

1. J. Denzler, C. M. Brown, and H. Niemann, "Optimal Camera Parameter Selection for State Estimation with Applications in Object Rrecognition," *Pattern Recognition. Proc. 23rd DAGM Symposium, München, Germany, 2001*. Ed. by. In B. Radig and S. Florczyk,. (Springer, Berlin Heidelberg, 2001), pp. 305-312.
2. R. I. Hartley, "In Defense of the Eight Point Algorithm," IEEE Trans. on Pattern Analysis and Machine Intelligence **19**, 580–593 (1997).
3. B. Heigl, PhD Thesis (Univ. Erlangen-Nuremberg, Erlangen, 2003).
4. M. Levoy and P. Hanrahan, "Light Field Rendering," in *Proc. SIGGRAPH, New Orleans, USA, 1996* (ACM Press), pp. 31-42.
5. I. Scholz, J. Denzler, and H. Niemann, "Calibration of Real Scenes for the Reconstruction of Dynamic Light Fields," in *Proc. of lAPR Workshop on Machine Vision Application, Nara, Japan, 2002*, pp. 32-35.
6. B. Triggs, "Autocalibration and the Absolute Quadric," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, (IEEE Computer Society Press, 1997), pp. 609–614.
7. F. Vogt, C. Klimowicz, D. Paulus, *et al*, "Bildverarbeitung in der Endoskopie des Bauchraums," in *5. Workshop Bildverarbeitung für die Medizin, Lübeck, 2001*, Ed by H. Handels, A. Horsch, T. Lehmann, and H.-P. Meinzer (Springer, Berlin, Heidelberg), pp. 320-324.
8. F. Vogt, D. Paulus, B. Heigl, *et al*, "Making the Invisible Visible: Highlight Substitution by Color Light Fields, in *Proc. First European Conf. on Colour in Graphics, Imaging, and Vision, Poitiers, France, 2002* (The Society for Imaging Science and Technology, Springfield, USA), pp. 352-357.
9. J. Borenstein, H. R. Everett, and L. Feng, *Navigating Mobile Robots – Systems and Techniques* (A. K. Peters, Wellesley, Massachusetts, USA, 1996).
10. D. Kortenkamp, A. P. Bonasso, and R. Murphy, *Artificial Intelligence and Mobile Systems* (AAAI Press, MIT Press, Cambridge, Massachusetts, 1998).
11. A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice* (Springer, Berlin, 2001).
12. F. Vogt, S. Krüger, H. Niemann, and C. H. Schick, "A System for Real-Time Endoscopic Image Enhancement" in (Editors): *Medical Image Computing and Computer Assisted Intervention (MICCM), Montreal, Canada 2003*, Ed. by R. E. Ellis, T. M. Peters (Springer, Berlin, 2003), Lecture Notes in Computer Science (LNCS), pp. 356–363.
13. P. Besl and N. McKay, "A Method for Registration of 3D Shapes," IEEE Trans. on Pattern Analysis and Machine Intelligence **14** (2), pp. 239–256 (1992).

Heinrich Niemann. Born 1940. Obtained his degree of Dipl.-Ing. in Electrical Engineering and Dr.-Ing. from Technical University of Hannover, Germany, in 1966 and 1969, respectively. Worked at Fraunhofer Institut für Informationsverarbeitung in Technik und Biologie, Karlsruhe, and at Fachhochschule Giessen in the Department of Electrical Engineering. Since 1975, Professor of Computer Science at the University of Erlangen-Nuremberg and the Dean of the Engineering Faculty of the university in 1979–1981. Since 1988, the Head of the research group "Knowledge Processing" at the Bavarian Research Institute for Knowledge-Based Systems (FORWISS). Since 1998, a speaker of a special research area (SFB) "Model-Based Analysis and Visualization of Complex Scenes and Sensor Data" finantially supported by the German Research Foundation (DFG). Scientific interests: speech and image understanding and the application of artificial intelligence techniques in these fields. Member of the editorial boards of *Signal Processing, Pattern Recognition Letters, Pattern Recognition and Image Analysis*, and *Journal of Computing and Information Technology.* Author and coauthor of seven books and about 400 journal and conference contributions. Editor and coeditor of 24 proceeding volumes and special issues. Member of DAGM, ISCA, EURASIP, GI, IEEE, and VDE and a Fellow of IAPR.

Joachim Denzler. Born 1967. Obtained his Diplom-Informatiker and Dr.-Ing. degrees from the University of Erlangen in 1992 and 1997, respectively. Currently, Professor of computer science and heads the computer vision group, Faculty of Mathematics and Computing, University of Passau. His research interests comprise active computer vision, object recognition and tracking, 3D reconstruction and plenoptic modeling, and computer vision for autonomous systems. Author and coauthor of over 70 journal papers and technical articles. Member of the IEEE, IEEE computer society, and GI. For his work on object tracking, plenoptic modeling, and active object recognition and state estimation, was awarded with DAGM best paper awards in 1996, 1999, and 2001, respectively.

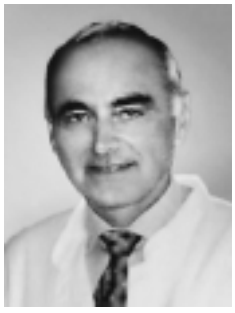
Benno Heigl. Born 1972. received his Diplom-Informatiker degree in Computer Science from the University Erlangen-Nuremberg, Germany, in 1996. Joined the Institute for Pattern Recognition until March 2000. Currently, works at Siemens Medical Solutions, Forchheim, Germany, dealing with 3D reconstruction from angiographic X-ray images and with image acquisition systems. His research interests include camera calibration, reconstruction of camera motion from monocular image sequences, and plenoptic scene modeling. Author and coauthor of over 20 publications



Florian Vogt. Born 1975. Obtain his Diplom-Informatiker degree from the University of Ulm in 2000. Currently, PhD fellow at the Chair for Pattern Recognition, University of Erlangen-Nuremberg. Scientific interests: computer-assisted endoscopy, endoscopic image enhancement, light field visualization for minimal-invasive operations. Author or coauthor of 15 papers.

Christoph H. Schick. Born 1962. Graduated from the Friedrich-Alexander-University, Erlangen-Nuremberg, in 1988 and got his PhD in 1991. Since 2003, Assistant Professor at the Department of Surgery, University Hospital, Erlangen. His areas of research are: computer-assisted surgery, image enhancement, robotics, integrated OR-systems, surgical oncology, and sympathetic surgery. Author of 31 papers. Member of "Deutsche Gesellschaft für Chirurgie," "Deutsche Gesellschaft für Endoskopie und Bildgebende Verfahren," and "International Society of Sympathetic Surgery." Board member of the International Society of Sympathetic Surgery (ISSS).

Sophie M. Krüger. Born 1973. Graduated from the University of Regensburg in 1999 and got her PhD in 2000. Currently, works at the Department of Surgery, University Hospital, Erlangen, as scientific assistant at the Sonderforschungsbereich 603 of the DFG. Member of "Deutsche Gesellschaft für Chirurgie." Her areas of research are: computer assisted surgery and image enhancement. Author of 10 papers.



Werner Hohenberger. Born 1948. Graduated from the Friedrich-Alexander-University, Erlangen-Nuremberg in 1974. Obtained his PhD in 1973. Since 1988, Professor at the University of Erlangen-Nuremberg, since 1995, Chairman of the Department of Surgery. His areas of research are: surgical oncology, coloproctology, and sepsis. Author of over 200 articles. Member of "Bayerische Krebsgesellschaft," "Deutsche Krebsgesellschaft," "Deutsche Gesellschaft für Chirurgie," "European Association of Coloproctology," "Vereinigung der Bayerischen Chirurgen," "Vereinigung der Gastroenterologen in Bayern," and "Deutschen Krebshilfe." Member of editorial boards of several journals, e.g., *Colorectal Disease*, *Digestive Surgery*, and *Langenbeck's Archives of Surgery.*

SPELL: 1. Erlangen-NЯremberg, 2. handheld, 3. freeform, 4. odometry, 5. online, 6. situs, 7. haptic, 8. preoperative, 9. specularities, 10. endoscope, 11. AESOP, 12. superpositioning, 13. semi-automatic, 14. Rrecognition, 15. finantially