Making the Invisible Visible: Highlight Substitution by Color Light Fields

Florian Vogt[1], *Dietrich Paulus*[1,*], *Benno Heigl*[3], *Christian Vogelgsang*[2], *Heinrich Niemann*[1], *Günther Greiner*[2], *Christoph Schick*[4]
[1]*Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg*
Martensstr. 3, 91058 Erlangen, Germany
[2]*Lehrstuhl für Graphische Datenverarbeitung, Universität Erlangen-Nürnberg*
Am Weichselgarten 9, 91058 Erlangen, Germany
[3]*Siemens Medical Solutions*
Siemensstraße 1, 91301 Forchheim, Germany
[4]*Chirurgische Universitätsklinik, Universität Erlangen-Nürnberg*
Krankenhausstr. 12, 91054 Erlangen, Germany

Abstract

In this contribution we present a new technique of highlight substitution. From a color image sequence, acquired with a hand-held camera, a so-called *light field* is generated. Additionally, a highlight mask is calculated for each image of the sequence. The highlight mask is then used as a confidence map for the light field. This results in color pixel interpolations at highlight pixels, taken from images in which these pixels were not over-imposed by highlights, resulting in better images. We demonstrate the technique on medical endoscopic images and evaluate the results on both, natural and synthetic data.

1. Introduction

When recording color image sequences of natural scenes, highlights due to specular reflection may considerably disturb the observer. This is particularly the case when medical images are recorded and humid tissue is subject to inspection. For endoscopic images the problem even increases as light source and viewing direction are almost identical; thereby, surfaces orthogonal to the viewing direction are often over-imposed to such an extent, that the physicians can only guess the tissue at that position.

In this contribution we show how highlights – as well as other image degradations – can be removed from image sequences when a light field is created first, that is subsequently used to enhance image quality at locations, where the input images show defects. The light field is a four-dimensional structure for rendering virtual color images from arbitrary positions within a certain volume.

This work was funded by Deutsche Forschungsgemeinschaft (DFG) under grants SFB 603/TP B6 and SFB 603/TP C2. Only the authors are responsible for the content.

*Authors current address: University Koblenz-Landau, Institut für Computervisualistik, Rammsweg 1, 56016 Koblenz

We discuss methods of highlight detection in Sect. 2. The idea and theory of light fields (as in [7]) is introduced in Sect. 3; we extend the light field structure in Sect. 4 by confidence maps that allow us to integrate the results of highlight detection into the reconstruction of an image. We evaluate the algorithm in Sect. 5. In Sect. 6 we summarize all facets of color image processing of the proposed system and propose further aspects of our work.

2. Highlight Detection

For di-electric inhomogeneous material, a model for separating specular reflectance from diffuse reflection exists [12], the so-called di-chromatic reflectance model. Algorithms based on this model have been applied, e.g., to remove highlights for stereo vision [8].

In [4], color gradients are used to detect highlights. Based on the RGB values, two new color spaces, $c_1c_2c_3$ and $l_1l_2l_3$, are defined [3] (assuming the di-chromatic reflectance model) and the color gradients are calculated in the three color spaces RGB, $c_1c_2c_3$ and $l_1l_2l_3$. The new color spaces are defined such that highlight edges are detected in the RGB and $c_1c_2c_3$ color spaces, but not in the $l_1l_2l_3$ color space. The disadvantage of this approach is that only edges are detected. To obtain highlight regions, some post-processing has to be done.

Human tissue, however, does not fit the model of di-electric inhomogeneous material. Experiments show, that in some cases reasonable results can still be obtained using this (incorrect) assumption for skin, e.g. in [10]. In our experiments, this assumption leads to poor detection of highlights in endoscopic images taken from the abdominal cavity. The reasons may be, that illumination is turned to red color by inter-reflections of red mucosa.

Assuming that no over-imposure is present in the images, highlights are simply detected in the HSV color

space by thresholds on the *saturation* and *value*. The obtained highlight mask is dilated (3×3 window) to obtain closed highlight regions. This also detects white colored areas as highlights; but as such areas are not present in the abdomen, this problem does not occur in endoscopic images taken from there.

Results for both methods are shown in Figure 1 for an endoscopic image of the thorax. In the following we use simple thresholds in HSV.

3. Light Field Generation and Visualization

Light fields have recently been introduced into computer vision and graphics [5, 9] and in general describe a sampled set of the plenoptic function [1] which is suitable for the generation of novel views of a scene.

The plenoptic function

$$\Phi = \Phi(x, y, z, t, \omega, \phi, \lambda) \tag{1}$$

measures the outgoing radiance of a specific wavelength λ at every point in space $(x, y, z)^{\mathrm{T}}$ in any direction $(\omega, \phi)^{\mathrm{T}}$ at any point in time t. This high dimensional space is not practical and thus the light field reduces the complexity to four dimensions by only measuring constant radiance along rays and by storing three channel color data. A ray is described by its intersection points (s, t) and (u, v) on two parallel planes. By setting up more slabs of two planes in space, a wide range of directions is covered. The rays starting from a fixed point on the st-plane and passing through all samples on the uv-plane describe all pixels of an image taken at the fixed location. Therefore a light field can be seen as a set of images captured at fixed locations and all of them sharing the same image plane.

A major challenge for computer vision, image processing, and graphics is to construct a light field from a real world scene. First an image sequence is captured by an uncalibrated camera moved on an unknown trajectory. Then this sequence is calibrated.

In this application the task of the calibration is the determination of camera motion from a continuously recorded image stream. Because of the low resolution of input images and their low signal-to-noise ratio an approach is required that is highly robust against mismatches and inaccuracies of matched image parts.

Most approaches are based on the knowledge of point correspondences between neighboring views which are determined for feature windows that can be tracked in arbitrary directions. As we have recorded a continuous image stream a corresponding feature in one frame appears close to the corresponding location in the previous frame. This property is exploited by differential tracking approaches as for example [13] being applied here.

Knowing corresponding point features, a lot of mathematical methods exist for computing camera motion from projections assuming a rigid scene [6]. The algorithms can be divided into those that are based on knowledge about intrinsic camera parameters (e.g. focal length) and those that also estimate these parameters. Because of increasing robustness we chose an approach of the first class in this scenario for reducing the total number of estimated parameters. The intrinsic camera parameters are determined in advance using a calibration pattern and applying the method [15].

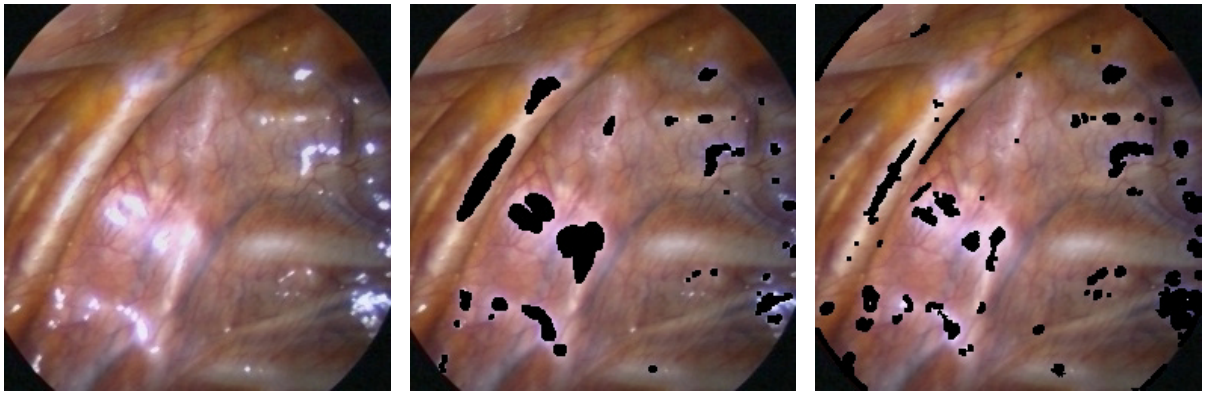
Our experiments have shown that the weak-perspective version [11] of the originally orthographic factorization method [14] is very robust against occurring outliers. The bias caused by the weak-perspective approximation is reduced by a following non-linear optimization step that is based on the (true) perspective projection model. As a side product of factorization, for each point that could be tracked over some frames the corresponding 3-D point is reconstructed. The result is a sparse geometrical representation of scene surface. It is interpolated yielding approximative dense depth maps that are used for visualization.

Finally a calibrated image stream is available and each frame contains an image and an associated viewing frustum describing the camera parameters. Unfortunately the cameras do not lie on the required two-plane setup and an additional warping step is required. During the warping, images of new, virtual cameras (target cameras) lying on the st-plane are generated: all pixels of contributing source cameras (the k nearest cameras are used; k is a parameter and usually $k = 5$ is sufficient) are projected into the 3D space and are then reprojected into the image plane of the new camera. If more than one color value is assigned to a 3D point, the resulting value is interpolated. 3D points lying behind other points (seen from the target camera) are omitted, i.e. they do not contribute to the new pixel value. The set of target camera images is then used for visualization.

Light field rendering uses the sampled data set and reconstructs new views of the scene by interpolating the stored samples. For two-plane parameterized light fields exists a hardware-accelerated rendering approach which allows the reconstruction of novel views with interactive rates. Visually convincing results can only be achieved by using a dense sample set with a huge number of images. The Lumigraph [5] extends a light field by adding geometric data about the objects placed in the scene. In our case the approximative dense depth maps, which are a result of the calibration (factorization) step, are used as geometric data. With this additional information the visual quality of the reconstructed views can be increased drastically.

4. Highlight Substitution

As block matching is usually not possible at image borders, no points can be tracked there and therefore no 3D points are reconstructed. To discriminate between those pixels where no reliable 3D information is available and the others, a so-called *confidence map* was introduced.



(a) (b) (c)

Figure 1: *(a) Original image (b) highlights with HSV thresholds (c) highlights with di-chromatic reflectance model, color gradients and subsequent filling of closed contours.*

The computed confidence maps for the depth value have the same size as the input image; they are computed for each image. The confidence map is set to zero if no 3D information is available and to > 0 (currently always set to 1) if interpolated or real 3D information is available.

For rendering, the effect of a low confidence (currently zero) in the map is as follows: during warping, pixels with confidence value zero are projected to infinity and are therefore not used for interpolation. To avoid black pixels, i.e. none of the k nearest cameras can be used for interpolation, k has to be increased, in the worst case to the maximum number of cameras available. In our experiments we set k to the maximum number of cameras available.

The confidence map is now used for another purpose as well; we mask out highlighted regions in the intensity images, as we set the confidence values for the depth maps in these locations to zero.

When a confidence value is zero, in some cases the rendered value has to be interpolated from neighboring object points. In many cases, the estimated value will depend on neighbors that are very close on the object, resulting from pixels that were visible in other views. It may even happen that exactly this object point was visible in another view and that the estimated intensity and color will thus be replaced by the *real* value, without any interpolation.

The result is a light field that contains images in which highlight pixels are substituted (interpolated).

5. Experiments and Evaluation

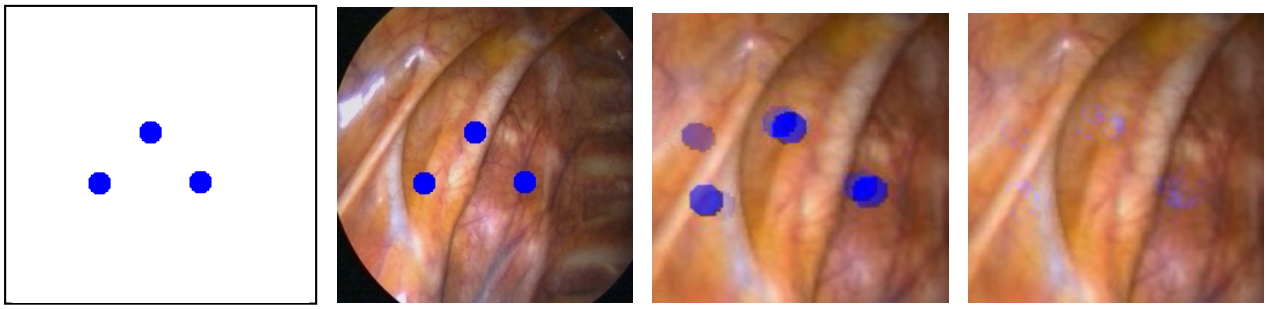
We evaluate the algorithm on synthetic and real image sequences. For synthetic images, we compute signal-to-noise ratios for an object with diffuse reflection in comparison to specular reflection. For real image sequences, we use endoscopic images that are evaluated in a double blind setup by medical physician.

As a first test and proof of the concept, circular spots (20 pixels diameter) have been colored (blue) by a mask in each image (size 256×256) of the input sequence as shown in Figure 2. We set the confidence map to zero at these locations. A light field has then been computed from these images and maps. Figure 2c shows a rendered image from the light field without using the mask as confidence map, Figure 2d shows a rendered image from the light field using the mask as confidence map. As can be seen, the blind spots are filled almost completely by information that has been taken from views that showed these areas when they were not masked out.

Next, we rendered two times 320 synthetical images from a scene containing a sphere and a cylinder, both with color texture (red and blue chessboard). One set of the synthetic images was rendered with highlight reflections and the other set without. An example with highlights is shown in figure Figure 3a. The reconstruction and calibration was done using the highlight data set. Using this reconstruction, three light fields were generated: one light field using the data set without highlights for rendering (ground truth G), one light field using the data set with highlights for rendering (L), and one light field using the data set with highlights *and* confidence map (*HSV*-threshold) for rendering (LC). Examples of rendered images of the light fields L and LC are shown in Figures 3b and 3c.

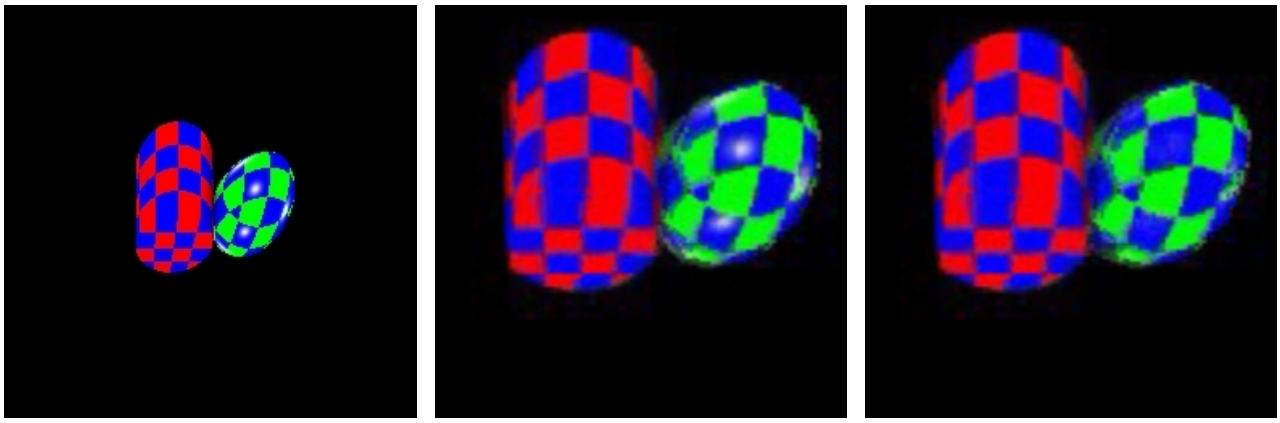
We rendered 50 images from each light field at the same position with the same camera parameters (randomly chosen) and calculated the mean signal-to-noise ratio between the images of the light fields L and G and the images of the light fields LC and G. Let a color image $\boldsymbol{f}$ with N rows and M columns be defined as

$$\boldsymbol{f} = \left[\boldsymbol{f}_{ij}\right]_{i=0,\ldots,N-1,j=0,\ldots,M-1} \tag{2}$$



(a) (b) (c) (d)

Figure 2: *(a) mask (b) mask overlaid over an original image (c) reconstructed view from light field without using the mask as confidence map (d) reconstructed view from light field using the mask as confidence map.*



(a) (b) (c)

Figure 3: *Example images of the synthetic sequence: (a) original image (b) image rendered from the light field without using the highlight mask as confidence map (c) same image as (b) except the highlight mask was used as confidence map.*

where

$$\boldsymbol{f}_{ij} = (r_{ij}, g_{ij}, b_{ij})^{\mathrm{T}} \ . \tag{3}$$

The mean signal-to-noise ratio was calculated as

$$\overline{SNR} = \frac{1}{50} \sum_{k=1}^{50} \frac{\sum_{i,j} \sum_{s \in \{r,g,b\}} s_{ij}}{\sum_{i,j} \sum_{s,n \in \{r,g,b\}} |s_{ij} - n_{ij}|} \tag{4}$$

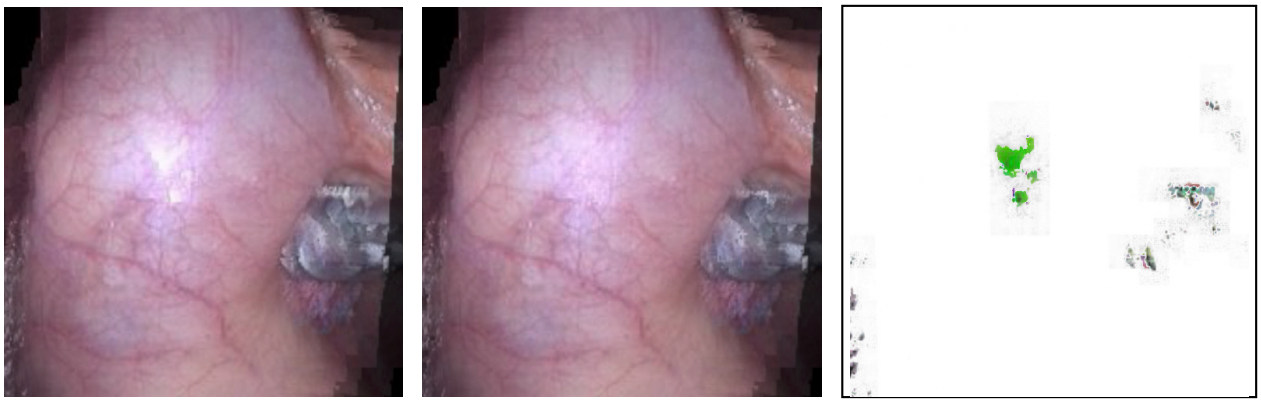
where the signal s was defined as the image rendered from G and the noise $|s - n|$ was defined as absolute value of the image difference between the signal image s and the noisy image n.

The mean signal-to-noise ratio ($\pm$ standard deviation) between L and G was 8.30 (± 0.49) and the mean signal-to-noise ratio between LC and G was 8.84 (± 0.83).

Medical light fields were generated from two endoscopic sequences: a sequence of the gall and a sequence of the thoracic cavity. The rigidity of a medical scene is not necessarily guaranteed because of respiration and heart activity. This fact has been ignored exploiting the fact that the chosen calibration approach is robust against small changes of detected point features.

Highlights were detected in the gall sequence in each image by *HSV* thresholds. The possible values for *H*, *S* and *V* were $H \in [0, 359]$, $S \in [0, 255]$ and $V \in [0, 255]$. We used the following thresholds: $0 \leq H \leq 359$, $0 \leq S \leq 20$ and $0 \leq V \leq 200$. Afterwards the binary highlight mask was dilated three times. The thresholds for the thoracic cavity were: $0 \leq H \leq 359$, $0 \leq S \leq 40$ and $0 \leq V \leq 200$ and the binary highlight mask was dilated only two times. Figures 4 and 5 show two selected images. The images were rendered from the lights fields and show the reduction of highlights. The difference images (Figures 4c and 5c) clarify the effect.

In a double blind setup, medical physicians evaluated rendered images from the gall and the thoracic cavity light field, at each case 50 images were rendered without using the highlight mask and 50 images (at the same positions with the same camera parameters) were rendered using



(a) (b) (c)

Figure 4: *Rendered images of the gall light fields at the same camera position with the same camera parameters: (a) light field without using the highlight mask as confidence map (b) light field using the highlight mask as confidence map (c) difference image* $|(a) - (b)|$*, inverted and rescaled: pixel values* > 100 *were set to* 255 *and pixel values* $\in [0, 100]$ *were transformed linearly to* $[0, 255]$.

the highlight mask. These two times 50 images were compared pairwise. The physicians selected almost *always* the image where the highlights were reduced to be the image with the higher or better quality: 45 of 50 images at the gall sequence and 50 of 50 images at the thoracic cavity sequence.

6. Conclusion

Using light fields even invisible, disturbed, or missing information in one view can be replaced by its *real* value, provided that this area on the object is visible in another view of the image sequence. Rather than substituting the information by heuristics, we use real information. This will work as long as the objects and the lighting conditions remain almost unchanged during recording of the sequence.

We showed that fixed image degradations can be removed by setting the confidence map of the generated light field at these locations to zero (cf. Figure 2). Fixed image degradations do not move when the camera moves, e.g. particles lying on the lense of the camera. In our experiments the (synthetic) degradations used were colored circles with 20 pixels diameter overlaid over the original images of size 256×256.

For synthetic data, we showed that the signal-to-noise ratio could be increased substantially considering that the number of highlight pixels is less than 10% of the object's area by substituting highlights. For real data the evaluation results of the physicians demonstrate that the image quality of the endoscopic images was enhanced by the highlight substitution.

We showed, how color image sequences can be enhanced by combining various strategies in computer vision and image processing. In the summary, we emphasize again the relation of the work to color processing:

- color image sequences are enhanced,
- highlights are detected by color imaging methods,
- color light fields are generated and used for highlight substitution,
- color interpolation is required for intensities at yet unknown pixel locations.

Our major application of this strategy is endoscopic medical imaging. Other applications of light fields in computer vision can be found, e.g. in [2], where light fields are used for self localization of a robot.

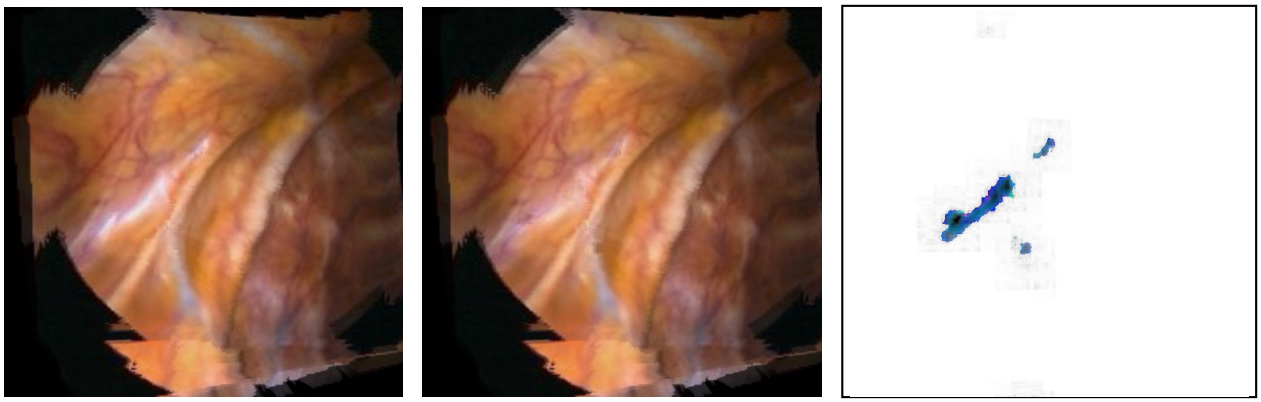
References

[1] E. H. Adelson and J. R. Bergen. *Computational Models of Visual Processing*, chapter 1 (The Plenoptic Function and the Elements of Early Vision). MIT Press, Cambridge, MA, 1991.

[2] J. Denzler, B. Heigl, and H. Niemann. Combining computer graphics and computer vision for probabilistic self–localization. *Machine Graphics & Vision*, page submitted, 2000.

[3] T. Gevers and A. W. M. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, 1999.

[4] Th. Gevers and H. M. G. Stokman. Classifying color transitions into shadow-geometry, illumination highlight or material edges. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages I:521–524, Vancouver, BC, September 2000. IEEE Computer Society Press.



(a) (b) (c)

Figure 5: *Rendered images of the thoracic cavity light fields at the same camera position with the same camera parameters: (a) light field without using the highlight mask as confidence map (b) light field using the highlight mask as confidence map (c) difference image* $|(a)-(b)|$*, inverted and rescaled: pixel values* > 100 *were set to* 255 *and pixel values* $\in [0, 100]$ *were transformed linearly to* $[0, 255]$.

[5] S. J. Gortler, R. Grzeszczuk, R. Szelinski, and M. F. Cohen. The lumigraph. *Computer Graphics (SIGGRAPH '96 Proceedings)*, pages 43–54, August 1996.

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[7] R. Koch, B. Heigl, M. Pollefeys, L. Van Gool, and H. Niemann. A geometric approach to lightfield calibration. In F. Solina and A. Leonardis, editors, *Computer Analysis of Images and Patterns — CAIP '99*, number 1689 in Lecture Notes in Computer Science, pages 596–603, Heidelberg, 1999. Springer.

[8] A. Koschan. Analyse von Glanzlichtern in Farbbildern. In D. Paulus and Th. Wagner, editors, *Dritter Workshop Farbbildverarbeitung*, pages 121–127 & 95, Stuttgart, 1997. IRB-Verlag.

[9] Marc Levoy and Pat Hanrahan. Light field rendering. In *Computer Graphics Proceedings, Annual Conference Series (Proc. SIGGRAPH '96)*, pages 31–42, 1996.

[10] C. Palm, T. Lehmann, and K. Spitzer. Bestimmung der Lichtquellenfarbe bei der Endoskopie makrotexturierter Oberflächen des Kehlkopfs. In K.-H. Franke, editor, *5. Workshop Farbbildverarbeitung*, pages 3–10, Ilmenau, 1999. Schriftenreihe des Zentrums für Bild- und Signalverarbeitung e.V. Ilmenau.

[11] C. J. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):206–218, March 1997.

[12] S. A. Shafer. Using color to separate reflection components. *COLOR research and application*, 10(4):210–218, 1985.

[13] J. Shi and C. Tomasi. Good features to track. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, Seattle, WA, June 1994. IEEE Computer Society Press.

[14] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, November 1992.

[15] R. Y. Tsai and T. S. Huang. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):13–27, 1984.