2006 IEEE International Conference on
Multisensor Fusion and Integration for Intelligent Systems
September 3-6, 2006, Heidelberg, Germany

Robust Depth Estimation by Fusion of Stereo and Focus Series Acquired with a Camera Array

Christian Frese and Ioana Gheţa

Abstract— **In order to obtain depth information from intensity sensors without auxiliary means, an image series is needed, gathered by varying at least the geometrical or focus position of the camera. Each of the fusion methods imposes certain constraints on the observed scene. With the aim of alleviating these restrictions, this contribution presents an algorithm to fuse combined stereo and focus series using energy functionals. A camera array is employed to record the series of images simultaneously.**

I. INTRODUCTION

A continuous challenge is the improvement of the capabilities of classical one camera systems. Often, scene properties can only be obtained by acquisition and fusion of multiple images. For example: a higher resolution can be obtained by the fusion of images with lower resolution, a higher dynamic by fusing images acquired with different exposure times, an authentic spectral information by fusing spectral series. Further applications include the generation of panorama images by fusion of image series acquired with different camera positions or high speed videos by the combination of different spatiotemporal images [1], [2], [3], [4].

In the field of visual inspection, depth maps play a major role. They describe for each image point the distance to the respective object in the scene. The most common methods are depth from focus, depth from defocus, depth from stereo, depth from shading and depth from motion, whereas for each of them a multitude of algorithms can be found in the literature. As the method names already denote, it is necessary, in each case, to acquire image series while changing the camera position or focus adjustment [5].

We propose a camera array system, that consists of several cameras arranged such that the optical axes are approximately parallel. The advantages lie in the degrees of freedom offered by such a system, i.e. each camera can be configured differently. Furthermore, the image series can be achieved with only one trigger, e.g. focus series are grabbed with the cameras having distinct focus positions. This leads to more robust results due to the possibility of applying and combining different algorithms. The capabilities of the camera array can be enhanced by positioning the entire system variably by a robot, e.g. for gaining panoramic depth images, Fig. 1.

Besides the hardware challenges of setting up a camera array, the presence of the stereo effect in the combined series also poses a difficulty to overcome. Because of the different camera positions, the images series are always combined ones, e.g. focus and stereo image series; with the exception, of course, when all cameras have identical focus parameters, which leads to pure stereo series. There are two possibilities for analyzing them: the first one is to evaluate and eliminate the stereo disparities and then fuse the remaining information (e.g. focus information). The second one is to fuse the images in the series under consideration of the different points of view. This paper concentrates on the second approach and presents an algorithm for fusion of combined stereo and focus series leading to a robust depth estimation.

Fig. 1. The camera array system in combination with an industrial robot

II. RELATED WORK

There have been some earlier attempts to fuse stereo and focus series for estimating depth. Bove [6] describes probabilistic models of depth from focus and stereo ranging, respectively. The fusion of range data is performed by weighted averaging using the local variance estimates as weights. These variance estimates are computed from the Cramer-Rao lower bound. However, Bove implemented a simple correlation based stereo algorithm. It seems difficult to carry out a similar statistical analysis for more sophisticated correspondence algorithms.

Krotkov [7], [8] also integrates stereo and depth from focus. He argues that both stereo correspondence and depth from focus may yield inconsistent measurement results and, hence, a weighted average fusion is not justified. Instead, Krotkov proposed a verification method: focus ranging is applied to eliminate false stereo matches. Similarly, depth from focus data may be verified by stereo ranging. Krotkov

C. Frese and I. Gheţa are with the Institute of Computer Science and Engineering, Chair for Interactive Real-Time Systems, Universität Karlsruhe (TH), Adenauerring 4, 76131 Karlsruhe, Germany, {frese; gheta}@ies.uni-karlsruhe.de

1-4244-0567-X/06/$20.00 ©2006 IEEE.

implements a feature based stereo algorithm which does not produce a dense depth map.

Subbarao et al. [9] integrate depth from defocus, depth from focus, and stereo ranging methods. A smaller search window for stereo correspondence analysis is determined using the depth from defocus and the depth from focus results. This reduces both the rate of matching failures and the computational efforts.

Rajagopalan et al. [10] propose a Markov random field approach for fusion of stereo and depth from defocus data. Depth cues and smoothness priors are integrated into an energy functional which is minimized using simulated annealing.

All described methods require multiple images from the same camera with varying parameters, whereas in our approach, the camera array takes only one image series with fixed parameter configuration. Therefore, the methods of [6], [9] are not applicable. Our strategy has some similarities to Krotkov's verification method, but we compute a dense depth map. As in [10], we develop an energy functional containing both stereo and focus information. Besides the fact that we use only one image series, the main differences are that we use focus measures instead of a depth from defocus approach and that energy is minimized using a more efficient graph cut algorithm instead of simulated annealing [11].

III. APPROACH

Stereo reconstruction can provide accurate depth estimates. However, severe errors may occur if the preceding correspondence algorithm establishes faulty pixel matches. Possible reasons for this problem are the presence of occlusions, low-structured objects, and matching ambiguities due to periodic textures. In this contribution, we develop methods to integrate focus information into the correspondence algorithm using energy functionals. In this way, we achieve a more robust correspondence analysis and thus a more reliable depth estimation. The article presents the fundamental ideas of depth from stereo and depth from focus regarding our algorithm, their combination and, in the end, the results obtained using image series acquired with the camera array.

IV. DEPTH FROM STEREO

A. Geometry and Reconstruction

1) Notation: We consider n gray-value images $g_1, \ldots, g_n$. To distinguish between pixel coordinates in different images, we attach the image index to the coordinate vector: $\mathbf{p} = (i, p_\mathrm{x}, p_\mathrm{y})^\mathrm{T}$, $i \in \{1, ..., n\}$. $g(\mathbf{p})$ describes the gray-value intensity of pixel $(p_\mathrm{x}, p_\mathrm{y})^\mathrm{T}$ in image g_i. The set of pixels $(i, p_\mathrm{x}, p_\mathrm{y})$ in image g_i is denoted by $\mathcal{P}_i$, and $\mathcal{P} := \mathcal{P}_1 \cup \ldots \cup \mathcal{P}_n$ is the set of all pixels. $\mathbf{p}$ and $\mathbf{q}$ are called corresponding pixels if they represent the same scene point in different views and are characterized by the symbol $\mathbf{p} \leftrightarrow \mathbf{q}$.

2) Epipolar Constraint: For any given pixel $\mathbf{p} \in \mathcal{P}_i$, it follows from the geometry of the imaging process that its corresponding pixel $\mathbf{q} \in \mathcal{P}_j$ must lie on a certain epipolar line within the image g_j (see Fig. 2). This can be described mathematically by the fundamental matrix $\mathbf{F}_{ji}$ [12]: for all corresponding pixels $\mathbf{p} = (i, p_\mathrm{x}, p_\mathrm{y})^\mathrm{T} \leftrightarrow \mathbf{q} = (j, q_\mathrm{x}, q_\mathrm{y})^\mathrm{T}$,

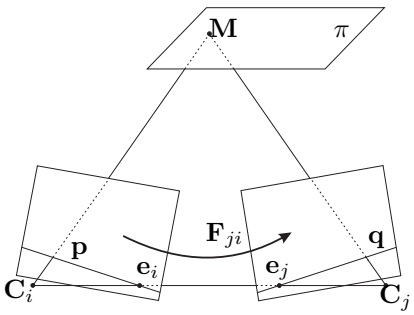

Fig. 2. $\mathbf{p}$ and $\mathbf{q}$ are projections of the same point in space $\mathbf{M}$ and are therefore correspondence pixels. Mathematically, their relation can be described with the help of the fundamental matrix $\mathbf{F}_{ji}$.

we have $(q_\mathrm{x}, q_\mathrm{y}, 1)\mathbf{F}_{ji}(p_\mathrm{x}, p_\mathrm{y}, 1)^\mathrm{T} = 0$ in homogeneous coordinates. The epipolar line of g_j corresponding to $\mathbf{p}$ is given by $\mathbf{F}_{ji}(p_\mathrm{x}, p_\mathrm{y}, 1)^\mathrm{T}$.

3) Calibration: The fundamental matrices are estimated from point correspondences using the normalized eight-point-algorithm [13]. The input point correspondences are obtained using a planar chessboard as calibration object.

We use a stratified reconstruction approach to calibrate the camera projection matrices [12]. Unlike the fundamental matrices calibration, in this case, it is necessary that the absolute 3D coordinates of the calibration points are known.

4) Reconstruction: Once the camera projection matrices are known, it is straightforward to reconstruct depth from point correspondences. This triangulation problem is equivalent to solving a system of linear equations [12].

5) Disparity Function: In binocular stereo, disparity is a function f that maps each pixel of the left image to its corresponding pixel in the right image. After applying an image rectification algorithm [14], the epipolar lines coincide with the image scanlines. Therefore, corresponding pixels differ only in their x-coordinates:

$$(1, p_\mathrm{x}, p_\mathrm{y})^\mathrm{T} \leftrightarrow (2, p_\mathrm{x} + f(1, p_\mathrm{x}, p_\mathrm{y}), p_\mathrm{y})^\mathrm{T}. \qquad (1)$$

This disparity function can be generalized for the camera array problem as follows: there is a disparity function $f : \mathcal{P} \to \mathcal{L}$, where $\mathcal{L}$ denotes the set of discrete disparity labels. We define the disparity $f(i, p_\mathrm{x}, p_\mathrm{y})$ to be α if and only if the pixel $(i, p_\mathrm{x}, p_\mathrm{y})^\mathrm{T}$ corresponds to a pair $(1, q_\mathrm{x}, q_\mathrm{y})^\mathrm{T} \leftrightarrow (2, q_\mathrm{x} + \alpha, q_\mathrm{y})^\mathrm{T}$. The abstract labels provide a generalization that describes point correspondences within the whole array, while the interpretation (1) remains valid for the reference pair consisting of cameras 1 and 2.

6) Point Transfer: Point transfer is the task of computing the coordinates of the corresponding pixel $(i, p_\mathrm{x}, p_\mathrm{y})^\mathrm{T}$, given the pixel $(j, q_\mathrm{x}, q_\mathrm{y})^\mathrm{T}$ and its disparity $\alpha = f(j, q_\mathrm{x}, q_\mathrm{y})$. If all camera centers are collinear, this can simply be accomplished by an addition to the x-coordinate, as shown above. In the general case, a projective homography is computed for each disparity label. The homography $\mathbf{T}_{ij\alpha}$ maps the coordinates of a pixel in g_j to the coordinates of the corresponding pixel

in g_i, provided that $f(j, q_\mathrm{x}, q_\mathrm{y}) = \alpha$:

$$(\tilde{p}_\mathrm{x}, \tilde{p}_\mathrm{y}, \tilde{p}_w)^\mathrm{T} = \mathbf{T}_{ij\alpha}(q_\mathrm{x}, q_\mathrm{y}, 1)^\mathrm{T}\,, \tag{2}$$

$$(i, p_\mathrm{x}, p_\mathrm{y})^\mathrm{T} = \left(i, \frac{\tilde{p}_\mathrm{x}}{\tilde{p}_w}, \frac{\tilde{p}_\mathrm{y}}{\tilde{p}_w}\right)^\mathrm{T}\,. \tag{3}$$

The 3×3 matrices $\mathbf{T}_{ij\alpha}$ can be computed as follows: select four non-collinear pixels in g_1. Addition of α to the x-coordinates yields the corresponding pixels in g_2. The epipolar lines in image g_i are computed using the fundamental matrices $\mathbf{F}_{i1}$ and $\mathbf{F}_{i2}$, respectively. Intersection of the resulting epipolar lines yields the corresponding pixels in g_i. In a similar manner, the corresponding pixels in g_j are computed. The homography $\mathbf{T}_{ij\alpha}$ is determined by these four corresponding pixels and its matrix can be computed using the inhomogeneous linear method, for example [12].

The transfer matrices $\mathbf{T}_{ij\alpha}$ are pre-computed. Consequently, point transfer can be accomplished by a single matrix-vector-multiplication.

B. Stereo Correspondence Formulation

We use the energy formulation of Kolmogorov and Zabih [15] to solve the stereo correspondence problem. They define the following energy functional, which depends on the disparity f:

$$E_\mathrm{stereo}(f) = E_\mathrm{data}(f) + E_\mathrm{smoothness}(f) + E_\mathrm{visibility}(f)\,. \tag{4}$$

The data term $E_\mathrm{data}(f)$ ensures photo-consistency, i.e. that corresponding pixels have similar gray-values. At first, a pixel dissimilarity measure d is needed. Suitable pixel dissimilarity measures include the squared intensity difference

$$d(\mathbf{p}, \mathbf{q}) = (g(\mathbf{p}) - g(\mathbf{q}))^2 \tag{5}$$

and the more sophisticated subpixel measure proposed by Birchfield and Tomasi [15], [16].

Using the dissimilarity measure, the data term can be defined as follows:

$$E_\mathrm{data}(f) = \sum_{(i,j)\in\mathcal{I}} \sum_{\substack{\mathbf{p}\in\mathcal{P}_i,\ \mathbf{q}=\mathbf{T}_{jif(\mathbf{p})}\mathbf{p} \\ f(\mathbf{p})=f(\mathbf{q})}} \min\{0,\ d(\mathbf{p}, \mathbf{q}) - K\}\,,$$

where $\mathcal{I}$ is the set of pairs of interacting images. It may be chosen as $\mathcal{I}_n = \{(i, j) :\ 1 \le i < j \le n\}$. For each pixel $\mathbf{p}$ of image g_i with assumed disparity $f(\mathbf{p})$, the corresponding pixel in image g_j is computed using the point transfer matrices derived above. The data term is relevant only if there is a pixel correspondence, i.e. if $\mathbf{p}$ and $\mathbf{q} = \mathbf{T}_{jif(\mathbf{p})}\mathbf{p}$ have equal disparities. The data term is always non-positive. Thus, each correct pixel correspondence decreases the total energy. The constant K defines an upper bound on pixel dissimilarity, which prevents outliers from influencing the energy minimization.

As the correspondence problem is underconstrained, additional knowledge is required to obtain a unique solution. A common assumption is that disparity should be piecewise constant. Discontinuities often coincide with intensity edges. These assumptions are modeled in the following smoothness term:

$$E_\mathrm{smoothness}(f) = \sum_{i=1}^{n} \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{N}_i} s(\mathbf{p}, \mathbf{q}, f)\,, \tag{6}$$

$$s(\mathbf{p}, \mathbf{q}, f) = \begin{cases} 0 & \text{if } \Delta f = 0 \\ \frac{1}{2}\lambda_1 & \text{if } \Delta f = 1 \text{ and } \Delta g < S \\ \frac{1}{2}\lambda_2 & \text{if } \Delta f = 1 \text{ and } \Delta g \ge S \\ \lambda_1 & \text{if } \Delta f > 1 \text{ and } \Delta g < S \\ \lambda_2 & \text{if } \Delta f > 1 \text{ and } \Delta g \ge S\,. \end{cases} \tag{7}$$

Herein, $\Delta f = |f(\mathbf{p}) - f(\mathbf{q})|$, $\Delta g = |g(\mathbf{p}) - g(\mathbf{q})|$, $\lambda_1 > \lambda_2$, $S > 0$ is a threshold for detecting intensity edges and $\mathcal{N}_i = \{(\mathbf{p}, \mathbf{q}) \in \mathcal{P}_i \times \mathcal{P}_i : |p_\mathrm{x} - q_\mathrm{x}| + |p_\mathrm{y} - q_\mathrm{y}| = 1\}$. Even though the discrete formulation of the smoothness term in [15] is shown to lead to good results, we have improved the behavior on slanted surfaces by adding the cases with $\Delta f = 1$.

The visibility term $E_\mathrm{visibility}(f)$ excludes some physically impossible disparity configurations by assigning them infinite energy. The reader is referred to [15] for more details on the visibility constraint and on the appropriate selection of the parameters S, λ_1, λ_2, and K. The terms λ_1 and λ_2 are scaled by a factor $\frac{2|\mathcal{I}|}{n}$ in order to maintain a constant ratio of data and smoothness terms independent of the number of images used.

The presented formulation has several advantages in the context of camera arrays and sensor fusion:

- Energy functionals allow to incorporate *a priori* knowledge explicitly. The terms $E_\mathrm{smoothness}$ and $E_\mathrm{visibility}$ in (4) model *a priori* assumptions about the disparity function.
- Further sources of information can easily be integrated by adding an appropriate term to the energy functional. We will make use of this possibility in Section VI to incorporate focus information.
- The formulation (4) is symmetric with respect to all cameras and thus the maximum of available information is exploited for depth estimation. The choice of the reference pair of cameras does not influence the resulting energy functional.
- The algorithm is state-of-the-art, as shown by its ranking in the evaluation by Scharstein and Szeliski [11].
- An effective minimization of the total energy is possible via graph cuts [17], [18], [19].

V. DEPTH FROM FOCUS

A. Image Formation Model

The literature provides a large number of algorithms that compute depth from focus, however by fusing image series acquired with only one camera [20], [21], i.e. from the same point of view. Focus series are "dense" image series, i.e. the changes in the focus settings are minimal from one image to the other. In the case of the camera array there are two problems to overcome: the number of focus positions is limited by the number of cameras in the array and the images in the series are taken from different points of view, such that the result is a combined stereo and focus series.

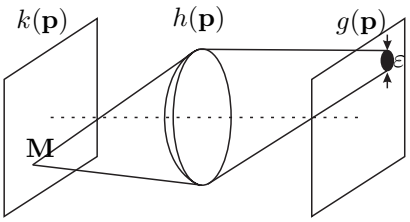

Fig. 3. The projection of the point $\mathbf{M}$ into the image plane takes the form of a blur circle with diameter ε.

The first problem can be solved by using more cameras. The second one is solved by the algorithm presented further on (see Section VI).

For a better understanding of the process, it has to be mentioned that the focus series are gathered with completely opened aperture. This reduces the depth of field and increases the blur of the object details that are out of focus. The blurring is modeled as a space variant convolution with a pulse response,

$$g(\mathbf{p}) = k(\mathbf{p}) * * \, h(\mathbf{p})\,, \tag{8}$$

$$h(\mathbf{p}) = \xi \text{rect}\left(\frac{||\mathbf{p}||}{\varepsilon}\right),\ ||\mathbf{p}|| = \sqrt{p_{\text{x}}^2 + p_{\text{y}}^2}\,, \tag{9}$$

$$\mathcal{F}\{h(\mathbf{p})\} := H(\mathbf{f}) = \frac{\pi\varepsilon^2}{4\xi}\frac{\mathbf{J}_1(\pi\varepsilon||\mathbf{f}||)}{||\mathbf{f}||}\,. \tag{10}$$

The last equation represents the Fourier transform of the pulse response, where $\mathbf{J}_1$ is the Bessel function of the first kind of first order, and ξ is a constant factor.

Fig. 3 shows the projection of a point in space to a blur circle on the image plane, where ε is the diameter of the blur circle.

The relation between the distance to the object and ε is the fundamental idea of depth from focus, i.e. the further away the object is, the bigger ε becomes.

B. Focus Measures

As defocused image formation is a low-pass operator (see (10)), the best focused image of an object contains the maximum amount of high frequency components. Therefore, it is common to use high-pass filters to compute feature images. For each pixel position, the feature image having the maximum focus measure represents the best focused view of the corresponding object. Depth is reconstructed from the known camera parameters of the respective image.

It is important to point out that focus measures are comparable only for images of a given object point. The absolute value of the focus measure strongly depends on the presence of object structures such as edges and texture. It is impossible to obtain any focus information from a single feature image. The maximization procedure is feasible only if all focus feature values considered correspond to the same object point.

The depth from focus method cannot be applied directly to our combined image series. The focus position of each camera is adjusted manually and remains fixed during the imaging process. Thus, we obtain a focus sequence in which each image has a different focus setting *and* a different perspective. Therefore, pixel correspondences have to be established before the maximization method can be applied to the image sequence. It follows that depth from focus is no longer independent from stereo correspondence analysis. Nevertheless, stereo correspondence analysis can be improved by integration of focus measures, as we will show in the next section.

VI. FUSION OF STEREO AND FOCUS SERIES

A. Verification of Disparity by Focus Measures

As mentioned above, the depth from focus method has to be modified in order to be applied to stereo and focus series. To this end, we may focus some cameras to different distances and compute the focus measures for the resulting images. Using disparity maps obtained from stereo, point correspondences can be established in the feature images. Then, it is possible to determine the maximum of the focus measure. This is used to verify the stereo disparity. The sketched algorithm is described in more detail subsequently.

Once the stereo disparities are known for some views, image warping techniques may predict the disparities in the other views [22]. At first, all pixels in the unknown disparity map $f_j := f(j,\cdot,\cdot)$ are labeled to be occluded. For each pixel $\mathbf{p}$ in the known disparity map f_i, the corresponding pixel $\mathbf{q}$ in g_j is computed using the point transfer matrices (see Section IV-A.6). The disparity value $f(\mathbf{q})$ is replaced by $f(\mathbf{p})$ if it previously has been occluded or if it has referred to an object farther from the camera. This incremental procedure avoids overwriting of nearer object points by farther ones.

In the next step, each pixel of the stereo disparity map is transferred to the focus feature images. If one or more of the warped disparities do not agree with the stereo disparity, the pixel is not considered further. In this case, it is likely that the correspondent 3D point is occluded in at least one of the cameras. Since the maximum of the focus measure might occur in the occluded view, we cannot rely on the depth from focus method in this case.

If there is no occlusion, the maximum of the focus measure is determined among the corresponding pixels. Together with the stereo disparity, we obtain two depth estimates. If both estimates are within the tolerance of measurement, the disparity is considered to be confirmed. Otherwise, at least one of the depth estimates must be wrong. Therefore, we label the depth estimate of the pixel under consideration as unreliable.

The following paragraph describes the precise criteria for deciding whether disparity and focus measure agree. The disparity corresponding to the plane of sharp focus is known for each camera. Therefore, the view with maximum focus measure can be predicted from the stereo disparity of a pixel. If it coincides with the actual focus maximum, the disparity receives the highest confidence possible. However, there is some uncertainty in the method because of calibration errors and the depth of field. Thus, the focus maximum may instead be detected in the view with the neighboring

plane of sharp focus, even if the disparity is correct. A lower confidence level is assigned to these disparity values. If the focus maximum is neither in the predicted view nor in the neighboring one, the disparity value is discarded. The three cases are illustrated in Fig. 4.

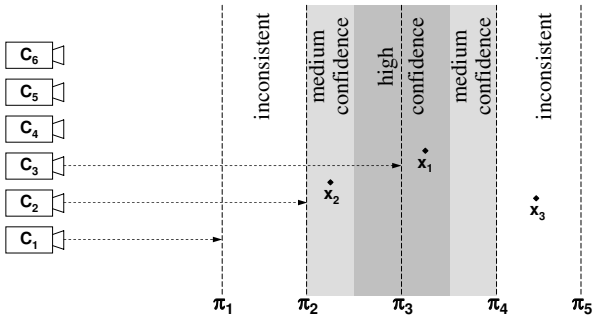

Fig. 4. The three confidence levels when fusing stereo and focus information. The cameras $\mathbf{C}_i$ are focused on the planes π_i. The confidence levels are shown for points which have maximum focus measure in view $\mathbf{C}_3$. Thus, $\mathbf{x}_1$ receives maximum confidence, as plane π_3 is nearest to the location of $\mathbf{x}_1$ reconstructed from stereo disparity. The depth estimate of $\mathbf{x}_2$ has medium level of confidence, as it is between π_3 and the neighboring plane π_2, but nearer to π_2. On the other hand, there is no confidence for the depth estimate of $\mathbf{x}_3$ because it is located outside the sector $[\pi_2, \pi_4]$.

B. Integration of Focus and Stereo Information

The method described in the previous section detects errors in disparity maps, but it cannot correct them. In order to obtain an improved disparity map, the focus measure has to be integrated in the stereo correspondence algorithm. This can be performed by adding another term to the energy formulation:[1]

$$E_{\text{fusion}}(f) := E_{\text{stereo}}(f) + E_{\text{focus}}(f)\,, \tag{11}$$

$$E_{\text{focus}}(f) := \sum_{\mathbf{p}\in\mathcal{P}} \begin{cases} 0 & \text{if } f(\mathbf{p}) \text{ has high confidence} \\ \mu_1 & \text{if } f(\mathbf{p}) \text{ has medium confidence} \\ \mu_2 & \text{if } f(\mathbf{p}) \text{ is inconsistent}\,, \end{cases}$$

where $\mu_2 > \mu_1 > 0$. The term $E_{\text{focus}}(f)$ is computed by transferring each pixel $\mathbf{p}$ of the stereo disparity maps to the focus feature images assuming a disparity of $f(\mathbf{p})$. The maximum of the focus measure is detected and one of the three confidence levels is assigned to the respective disparity value (cf. Fig. 4).

VII. EXPERIMENTAL RESULTS

This section presents some results of fusing stereo and focus series acquired with an experimental setup consisting of nine cost-effective cameras, arranged in form of a 3×3 matrix, so that the optical centers are coplanar. The cameras are all of the same type and equipped with the same type of optics, Fig. 1.

[1]A continuous form of E_{focus} is possible, e.g. by interpolation, however, the low number of focus positions is not appropriate for this approach.

A. Stereo

Using the multi-camera stereo algorithm without focus information, good results were obtained with many scenes. As expected, occlusion problems almost disappear with an increasing number of cameras. The slight differences in camera sensitivity are compensated by a linear gray-value transformation.

B. Integration of Focus Measures

Depth from stereo needs image features, e.g. edges or lines, for estimating correspondences. Therefore, in regions with low structure, faulty results can occur. Fig. 6 presents the effectiveness of the fusion of stereo and focus series for the scene of Fig. 5, which has low structure in the background. For a better visibility, only a fragment of the images is shown. The depth map in Fig. 6(b) contains severe errors up to 15 cm within the whole background because the observed object has little structure. This is confirmed by the confidence map in Fig. 6(c). When the inconsistent areas are eliminated from the depth map, Fig. 6(d) results. The depth map of Fig. 6(e) resulting from the evaluation of the series using the combined energy functional is correct almost within the entire background. However, there may be more errors in the foreground. This is improved when the depth maps in Fig. 6(b) and 6(e) are combined by replacing the depth values discarded by the verification procedure with data from the combined evaluation (see Fig. 6(f)).

Even though the energy functional model is extended by one term, the computational time required for the minimization does not increase significantly.

VIII. CONCLUSIONS AND FUTURE WORKS

For an efficient gaining and evaluation of combined stereo and focus series, a camera array is deployed. The algorithm for fusing the series combines depth from stereo and depth from focus. Starting from the known model for fusion of stereo image series based on energy functionals, a new enhanced fusion model that integrates depth from focus is

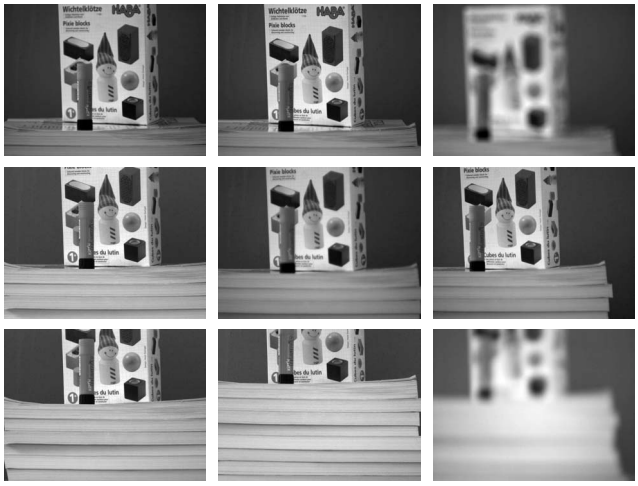

Fig. 5. Combined stereo and focus image series obtained with the 3 × 3 camera array. Each camera has an individual perspective and focus position.

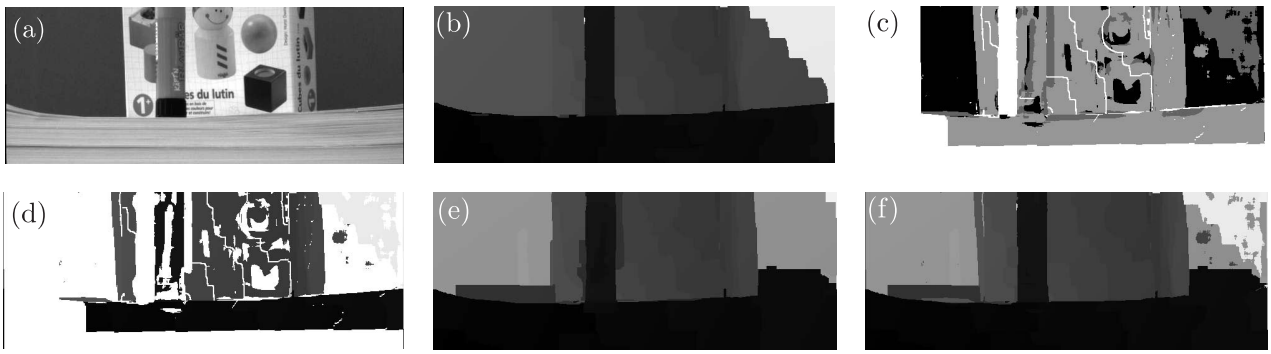

Fig. 6. Fusion of the image series in Fig. 5: (a) Fragment of an image with little texture (b) Depth map after pure stereo fusion: the depth of the background is erroneous (c) Confidence of the depth map, where black = inconsistent, gray = medium confidence, light gray = high confidence, and white = occlusions (d) Depth map without low confidence areas (e) Fusion of stereo and focus image series (f) Combination of (b) and (e)

developed. The practicability of the algorithm is shown for an exemplary scene.

Considering the potential of the camera array, there are several directions to be followed in future work. One is the extension of the depth maps to panoramic depth maps by using a robot to position the entire system or by enlarging the system with cameras featuring different directions of view.

Another interesting possibility is to fuse the stereo depth map with a sparse depth map, where the distance to the object is only measured on image edges. This approach takes into consideration the fact that focus information is more reliable on edges, so the sparse depth map is computed by fusing information obtained from blurred edges.

REFERENCES

[1] R. Szeliski, "Image mosaicing for tele-reality applications," in *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, 1994, pp. 44–53.

[2] Y. Schechner and S. Nayar, "Generalized mosaicing," in *Proceedings of the Eighth IEEE International Conference on Computer Vision*, 2001, pp. 17–24.

[3] P. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of SIGGRAPH*, 1997, pp. 369–378.

[4] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," *Proceedings of ACM SIGGRAPH*, vol. 24, no. 3, pp. 765–776, 2005.

[5] F. Puente León and J. Beyerer, "Datenfusion zur Gewinnung hochwertiger Bilder in der automatischen Sichtprüfung," *Automatisierungstechnik*, vol. 45, no. 10, pp. 480–489, 1997.

[6] V. M. J. Bove, "Probabilistic method for integrating multiple sources of range data," *Journal of the Optical Society of America A*, vol. 7, no. 12, pp. 2193–2198, 1990.

[7] E. P. Krotkov, *Active computer vision by cooperative focus and stereo*. Springer, 1989.

[8] "Active vision for reliable ranging: Cooperating focus, stereo, and vergence," *International Journal of Computer Vision*, vol. 11, no. 2, pp. 187–203, 1993.

[9] M. Subbarao, T. Yuan, and J.-K. Tyan, "Integration of defocus and focus aalysis with stereo for 3d shape recovery," in *Proceedings of The International Society for Optical Engineering (SPIE)*, vol. 3204, 1997, pp. 11–23.

[10] A. N. Rajagopalan, S. Chaudhuri, and U. Mudenagudi, "Depth estimation and image restoration using defocused stereo pairs." *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 11, pp. 1521–1525, 2004.

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms." *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.

[13] R. Hartley, "In defence of the 8-point algorithm." in *International Conference on Computer Vision*, 1995, pp. 1064–1070.

[14] C. T. Loop and Z. Zhang, "Computing rectifying homographies for stereo vision." in *Conference on Computer Vision and Pattern Recognition*, 1999, pp. 1125–1131.

[15] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts." in *European Conference on Computer Vision*, 2002, pp. 82–96.

[16] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling." *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 4, pp. 401–406, 1998.

[17] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 2, pp. 147–159, 2004.

[18] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision." *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 9, pp. 1124–1137, 2004.

[19] V. Kolmogorov, "Match. an implementation of three stereo algorithms and of the maxflow algorithm," 2003. [Online]. Available: http://www.cs.cornell.edu/~rdz/graphcuts.html

[20] S. K. Nayar and Y. Nakagawa, "Shape from focus." *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, no. 8, pp. 824–831, 1994.

[21] M. Subbarao and T. Choi, "Accurate recovery of three-dimensional shape from image focus." *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 3, pp. 266–274, 1995.

[22] M. Bleyer and M. Gelautz, "A layered stereo algorithm using image segmentation and global visibility constraints." in *Proc. of the 2004 International Conference on Image Processing*, 2004, pp. 2997–3000.