# Depth and all-in-focus images obtained by multi-line-scan light-field approach

Svorad Štolc[a,b] and Reinhold Huber-Mörk[a] and Branislav Holländer[a] and Daniel Soukup[a]

[a] AIT Austrian Institute of Technology GmbH, Austria
[b] Institute of Measurement Science, Slovak Academy of Sciences, Slovakia

## ABSTRACT

We present a light-field multi-line-scan image acquisition and processing system intended for the 2.5/3-D inspection of fine surface structures, such as small parts, security print, etc. in an industrial environment. The system consists of an area-scan camera, that allows for a small number of sensor lines to be extracted at high frame rates, and a mechanism for transporting the inspected object at a constant speed. During the acquisition, the object is moved orthogonally to the camera's optical axis as well as the orientation of the sensor lines. In each time step, a predefined subset of lines is read out from the sensor and stored. Afterward, by collecting all corresponding lines acquired over time, a 3-D light field is generated, which consists of multiple views of the object observed from different viewing angles while transported w.r.t. the acquisition device. This structure allows for the construction of so-called epipolar plane images (EPIs) and subsequent EPI-based analysis in order to achieve two main goals: (i) the reliable estimation of a dense depth model and (ii) the construction of an all-in-focus intensity image. Beside specifics of our hardware setup, we also provide a detailed description of algorithmic solutions for the mentioned tasks. Two alternative methods for EPI-based analysis are compared based on artificial and real-world data.

**Keywords:** light-field vision, line-scan acquisition, computational imaging, all-in-focus, depth sensing, surface analysis, industrial inspection

## 1. INTRODUCTION

Depth information from images is typically obtained using specific setups, such as time-of-flight sensors, configurations based on laser triangulation or multi-camera systems. At the same time, in order to observe a larger or varying depth of field, approaches to extend or adapt the focal range are of wide interest. In this paper, we address both issues related to industrial applications. We present a light-field multi-line-scan image acquisition system intended for the 2.5/3-D inspection of fine surface structures, such as small parts, security print, etc.

The system consists of an area-scan camera, that allows for a small number of sensor lines to be extracted at very high frame rates and a mechanism for transporting the inspected object at a constant controlled speed. In our experiments, we have employed the AVT BONITO CL-400C camera in combination with a conventional conveyor belt. During the acquisition, the object is moved orthogonally to (i) the camera's optical axis and (ii) the orientation of the sensor lines. In each time step, a region of interest consisting of several lines is read out from the sensor and stored. Afterward, by collecting all corresponding lines acquired over time (i.e., all 1st lines form one image, all 2nd lines form another image, etc.), a 3-D light field is produced consisting of multiple views of the object observed from different viewing angles w.r.t. the system optical axis. As the field of view of the camera is limited in the transport direction by the number of lines extracted from the sensor (in our case about 2°), the proposed device can be considered as an extremely narrow-baseline multi-view stereo system, which, nevertheless, provides an information-rich light field when compared with a single stereo pair.

Beside the detailed hardware description of the proposed system, we address the following two objectives regarding the processing of the obtained light-field data: (i) the reliable estimation of a dense depth model and (ii) the construction of an all-in-focus color intensity image with an improved signal-to-noise ratio. To achieve both goals, light field-based processing is performed in the so called epipolar plane image (EPI) domain.[1] For

Corresponding author: svorad.stolc@ait.ac.at

Image Processing: Machine Vision Applications VII, edited by Kurt S. Niel, Philip R. Bingham,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 9024, 902407 · © 2014 SPIE-IS&T
CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2042475

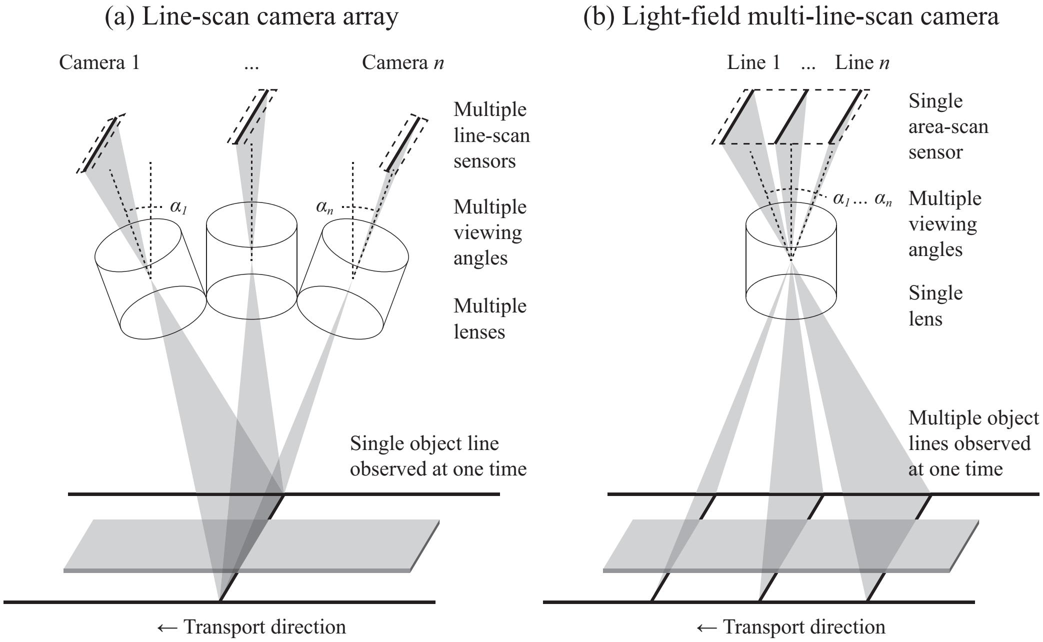


Figure 1. Equivalence between an array of ordinary line-scan cameras (a) and our multi-line-scan light-field camera (b).

both objectives, the algorithmic solutions are provided and the overall performance is demonstrated on synthetic light-field data as well as couple of real-world examples.

The paper has the following structure: Section 2 discusses technical details of our multi-line-scan light-field camera. At the beginning of Section 3, we give a brief overview of the state-of-the-art methods for a general light-field data processing. In Sections 3.1 and 3.2, we describe algorithms for the depth model estimation and the all-in-focus image construction from the obtained light-field data, respectively. In Section 4 we show results of our experiment with ground truth data and provide real-world examples of objects acquired by our acquisition setup and processed by our algorithms. Conclusions are drawn and discussed in Section 5.

## 2. MULTI-LINE-SCAN LIGHT-FIELD CAMERA

Light field capture and processing are methods of computational photography, an emerging field of research in computer vision. In general, computational photography is a combination of digital sensors, modern optics, actuators and smart lights to escape the limitations of traditional film cameras, enabling for novel imaging applications.[2] We restrict ourselves to acquisition and processing of light-field data in an industrial environment.

A light field is defined as the 4-D radiance function of 2-D position and 2-D direction in regions of space free from occluders.[3] Practical light field acquisition can be realized in various ways. The simplest and bulkiest approach is a multi-camera array using a number of cameras capturing the same scene at the same time from different viewing angles.[4] A gantry system uses a single camera which is mechanically displaced and a preferably still scene is acquired over time.[3] PiCam realizes a monolithic camera array using a 4 × 4 lens array placed on an equally organized sensor array on a single substrate.[5] An unstructured acquisition using a moving hand-held device was also described recently.[6] Splitting the optical path using filters, masks, code patterns, etc. is termed *coded aperture imaging*.[7] Recently, the use of microlens arrays placed in front of the senor plane was realized in plenoptic cameras, such as those manufactured by Lytro[8] and, finally, Raytrix[9] for industrial applications.

Our approach to the light field acquisition is tailored to the typical requirements of an industrial inspection task, namely we consider a line-scan camera acquisition setup. In our previous work,[10] we already explored the possibilities of multi-view line-scan systems using planar mirrors. This particular system can be considered a sparse light-field camera providing no more than 3 different views of the inspected object (i.e., 1 straight view and 2 tilted views).

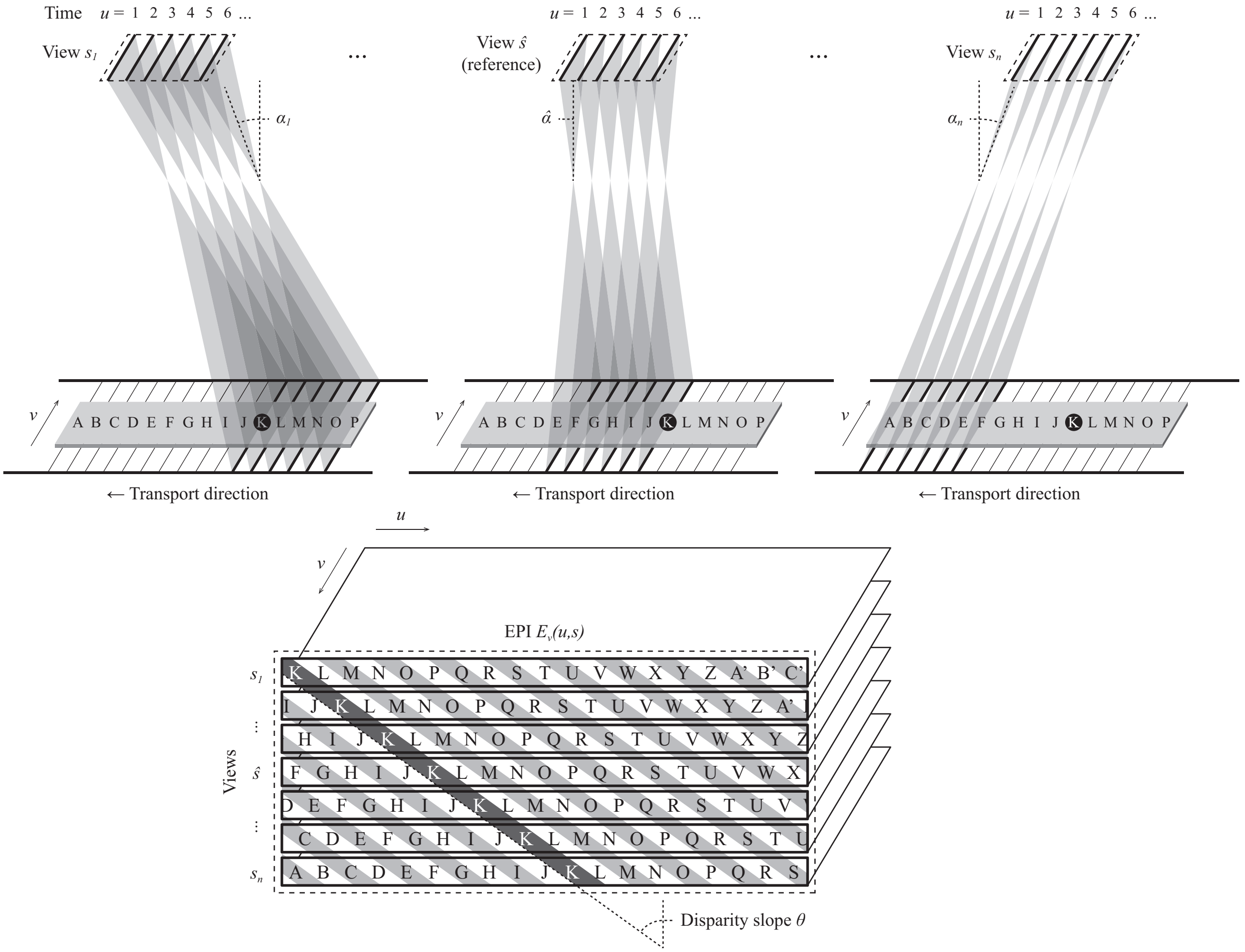

Time $u = 1\ 2\ 3\ 4\ 5\ 6 \dots$
View $s_1$
$\alpha_1$
View $\hat{s}$ (reference)
$\hat{\alpha}$
View $s_n$
$\alpha_n$
$v$
A B C D E F G H I J K L M N O P
← Transport direction
$u$
EPI $E_v(u,s)$
Views
$s_1$
$\hat{s}$
$s_n$
K L M N O P Q R S T U V W X Y Z A' B' C'
I J K L M N O P Q R S T U V W X Y Z A'
H I J K L M N O P Q R S T U V W X Y Z
F G H I J K L M N O P Q R S T U V W X
D E F G H I J K L M N O P Q R S T U V
C D E F G H I J K L M N O P Q R S T U
A B C D E F G H I J K L M N O P Q R S
Disparity slope $\theta$

Figure 2. Formation of multiple object views $s_i$ by the light-field multi-line-scan camera over time $u \in \{1, 2, 3, \cdots\}$ is outlined above. The obtained 3-D light-field data structure is shown below. The letters "A","B","C",... represent identifiable visual structures residing in the object plane, which serve us for demonstrating the effect of diagonal lines in the EPI domain. Note that slopes $\theta$ of individual EPI lines are proportional to the distance between the camera and the corresponding object point.

In our current approach, we extend this approach to provide a far greater amount of views, which in turn greatly enhances the quality of the depth reconstruction. In particular, an area-scan sensor is used to capture object positions $(u, v)$ under varying angle $s$ along the transport direction over time. Note that in this specific setup there is no variation of the angle $t$ measured across the transport direction. Therefore, our system acquires in fact only a 3-D slice of the complete 4-D light field. Fig. 1 shows the equivalence between an array of ordinary line-scan cameras observing a single object line from several angles at once and our light-field multi-line-scan camera, where a number of lines is observed from various angles at one time. In our setup, multiple views of a single object line are collected over time thanks to a relative motion between the camera and the object (see Fig. 2). In detail, Fig. 1 (a) depicts a multi-camera setup depicting an object at a single time moment from angular displaced line-scan cameras. Fig. 1 (b) shows an area-scan sensor observing the multiple object lines at one time. Due to motion, a specific object line is then observed under varying angles analogously to Fig. 1 (a).

In the following, we refer to an image formed by all object lines characterized by a constant viewing angle $s$ collected over time as a *view*. In computational imaging, a view is sometimes also called a *sub-aperture image*.

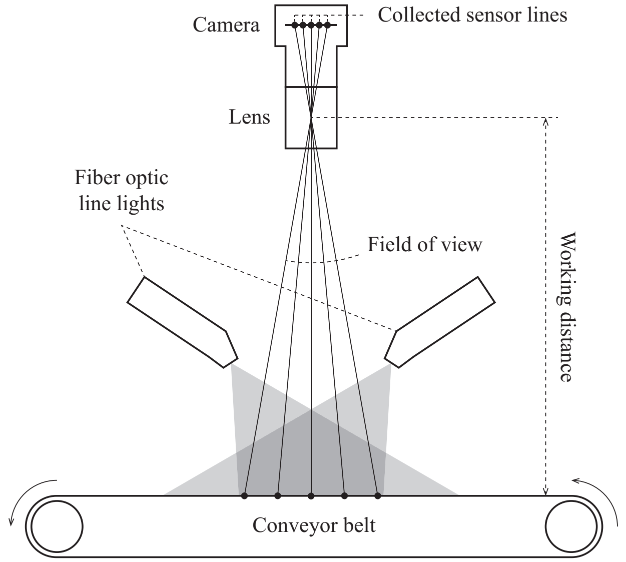
In Tab. 1, a schematic view of the suggested acquisition setup together with a detailed hardware specification is provided. Note that this kind of setup is quite typical in machine vision. Objects on the conveyor belt are imaged on by a single lens CMOS camera. The scene is illuminated by line lights, which in our case are not focused on a single line, as is usual in a line-scan imaging, but rather defocussed to be able to illuminate the whole (narrow) field of view.

Particular importance should be attributed to the choice of lenses. These determine the effective spatial resolution and hence the depth resolution achievable with the light-field system. One should also choose the lenses with geometrical distortions in mind (the individual lines which we acquire each frame should be parallel to each other). We consider two different lens configurations intended for lower and higher resolution applications, as described in Tab. 1 (below).

One can see that our system is actually a narrow-baseline multi-view stereo system, which means that the field of view is very narrow (approx. 2°, see Tab. 1, below). This significantly reduces the geometrical distortions and helps to guarantee that the obtained light-field data directly fulfills the epipolar constraint of stereo vision, which makes the post-processing significantly easier. E.g., given the non-canonical verged stereo geometry of our setup (i.e., the zero disparity is obtained in a finite distance from the camera), it can be shown that there is a simple linear relationship between the depth and detected disparity: $depth = disparity \times c_v$, where $c_v$ is a constant factor estimated independently for each data row $v$ in the calibration phase. Thus, for the sake of simplicity, in this paper, we restrict ourselves to the concept of disparity, which can, however, be easily converted to the actual depth value although requiring an additional calibration step. As already discussed, we use two different lenses which provides us with two different spatial and depth resolutions.

It should be noted that our light-field multi-line-scan camera is closely related to the *time delay and integration* (TDI) approach,[11] where the object is observed several times, but presumably under virtually the same angular

Table 1. Multi-line-scan light-field machine vision system.



| *Camera* | AVT BONITO CL-400C |
|---|---|
| – *Sensor* | Alexima AM41 CMOS area-scan |
| – *Resolution* | 2320 × 1726 px |
| – *Pixel pitch* | 7 $\mu$m |
| – *CFA* | Bayer pattern |
| – *Max. clock* | 200 MHz |
| – *Shutter* | global |
| – *Color res.* | 8 bit |
| – *Interface* | 2 x 10-tap Camera Link Full+ |
| – *Lens mount* | F-mount |
| *Illumination* | Volpi intraLED 5 with a dual branch fiber optic line light |
| *Conveyor belt* | Montech AG Minidrive KTB (Center drive) |

| | *Low-resolution option* | *High-resolution option* |
|---|---|---|
| *Lens* | NIKON AF NIKKOR 20 mm F2.8D | SIGMA EX DG Macro 50 mm F2.8 |
| *Focal length* | 20 mm | 50 mm |
| *Working distance* | 37 cm | 37 cm |
| *Spatial resolution* | 120 $\mu$m/px | 48 $\mu$m/px |
| *Field of view* | 128 lines (2.4°) | 256 lines (1.9°) |
| *Collected sensor lines* | 9 line pairs, one per each 16 lines* | 9 line pairs, one per each 32 lines* |
| *Stereo depth resolution* | 2.86 mm/unit disparity | 1.44 mm/unit disparity |

* Line pairs are necessary for full-color reconstruction on a Bayer FCA sensor.

directions. In our case, however, the object is observed each time from a slightly different viewing angle along the transport direction. We also note, that in a line-scan acquisition process, light fields can also be captured using additional optical elements, such as mirrors (e.g., Refs. 10).

## 3. DATA PROCESSING

The data corresponding to varying observation angles at each row of an area-scan sensor is shown in Fig. 2. E.g., the object region "K" in the focal plane is seen under different angles $s_i$ at different time instances $u$. Integration along a line with the slope $\theta$ provides the irradiance estimate for the in-focus object region "K". The data structure shown in Fig. 2 is called an *epipolar plane image* (EPI) when represented as a 2-D image. EPIs were originally introduced for estimation of structure from motion,[1] but they have also become a popular tool in light field processing (e.g., Refs. 12, 13).

It can be seen that parts of the object that do not reside in the focal plane of the camera result in a mapping to the sensor plane, which is different from the one described above. For such parts, corresponding radiance values map to lines as well, but with slopes reflecting a distance between camera and that particular object point. Consequently, the integration in EPIs along various slopes allows for refocusing to virtual focal planes different from the main focal plane of the camera. Furthermore, this effect can be used the other way round, i.e., by detecting predominant orientations of linear structures in EPIs, one can estimate the distance of the plane in which the corresponding object point would be in focus. A combination of both, the EPI-based local estimation of predominant orientations with the integration along the estimated slopes delivers the so-called *all-in-focus* image along with the dense depth estimate derived from estimated EPI slopes.

### 3.1 Disparity estimation

Naturally, there exist a number of strategies to determine predominant orientations in EPIs. One frequently used approach is a so-called slope hypotheses test. In this method a number of slope hypotheses in each location of the reference view (e.g., the central or straight vertical view) are verified. A winning hypothesis is identified according to some criteria reflecting how well the slope matches with the pattern in the EPI domain. A general approach to formulating such a criterion is to assume a Lambertian scene, which presumes that all object points preserve their radiance values regardless of the viewing angle. Furthermore, the illumination is supposed to remain constant for the different views in the light field. In the following, we refer to a scene meeting both assumptions as a *static Lambertian* scene. In general, natural environments with daylight ambient illumination are quite close to the static Lambertian scenes, especially when the stereo baseline is rather small w.r.t. the scene depth. However, in industrial applications, the static Lambertian assumption can be violated strongly due to various reasons, e.g., when observing glossy or specular materials like metals.

Kim et al.[13] build on the static Lambertian assumption by exploiting an easy criterion for ranking hypotheses, namely the best hypothesis is the one, for which as many radiance values as possible along the hypothesized slope in EPI are similar enough to the radiance in the reference view. In other words, the winning hypothesis minimizes the overall deviation of radiance values along its slope w.r.t. the reference.

Another approach, described by Venkataraman et al.,[5] uses pattern matching between different views. For a discrete number of hypothesized depths the *sum of absolute differences* (SAD) of radiances between different views is calculated. Again, this also dependents on the static Lambertian behavior to certain extent. The approaches by Kim et al. and Venkataraman et al. perform well in natural scenes, however, their performance might decrease under more general conditions.

Wanner and Goldlücke[14] suggest a statistical approach to estimate the principal orientation of linear structures in EPIs via analysis of the *structure tensor* constructed locally in small EPI neighborhoods. This method effectively eliminates the necessity of testing multiple slope hypotheses, which makes it well suited for high performance applications. However, it also assumes the static Lambertian behavior and, moreover, suffers from a rather high sensitivity to noise. To cope with this problem, the authors apply a global optimization (regularization) strategy, which significantly increases accuracy of resulting depth maps. However, the optimization introduces very high computational demands. Therefore, neither of the above mentioned methods provide an ultimate solution to challenges posed by industrial applications.

In this paper, we build on a well-known method of the disparity hypotheses testing for identification of predominant slopes in the EPI domain. For each location in the reference view (e.g., the central view), a number of disparity hypotheses are generated and the best one is taken for the final slope estimate. There exists a plethora of criteria how to assess hypotheses and identify a winning hypothesis. For very general conditions of industrial environments (i.e., materials with diverse surface properties ranging from matte to glossy, non-constant illumination from dark to bright field, etc.), we propose to use an extended version of the SAD-based method similar to Refs. 5, 13 – a so-called *modified SAD* (MSAD) – which provides significantly improved performance in such general conditions.

In order to prepare for the description of the MSAD method, let us first formulate the hypotheses testing algorithm based on SAD within the block matching framework frequently used in stereo vision. When applied to a stereo pair, block matching tries to find corresponding local image patches in both views and so estimate local disparities. To be able to apply the same approach to our light-field data, one has to consider an extension to block matching. In detail, image patches are compared between multiple views along hypothesized slopes in EPIs w.r.t. the reference view.

It should be noted that SAD strongly builds upon the static Lambertian assumption and, furthermore, it does not provide any special treatment for areas affected by object shadows or occlusions. Provided that the static Lambertian assumption is met, one may formulate a simple criterion for the best hypothesis making use of the SAD measure as follows: the winning hypothesis is the one, which minimizes the total SAD of all the blocks along the hypothesized EPI slope w.r.t. the corresponding block in the reference view. The SAD-based block matching can formally be described in the following way. Let us denote an $m \times m$ image patch associated with the object point $(u, v)$ observed from the angle $s$ as $\mathcal{P}_v^{m\times m}(u, s)$. Referring to the notion of EPIs, $\mathcal{P}_v^{m\times m}(u, s)$ is the image patch centered at the coordinate $(u, s)$ of $v$-th EPI, i.e., the patch extending from EPI orthogonally into the spatial image domain.

Further, we define a set of image patches $\Omega(u, v, \theta)$ collected along an imaginary line with a slope $\theta$, which intersects the reference view with an index $\hat{s}$ in a position $(u, v)$:

$$\Omega(u, v, \theta) = \left\{ \mathcal{P}_v^{m\times m}\left(u + \frac{s - \hat{s}}{\max(n - \hat{s},\ \hat{s} - 1)}\,\theta,\ s\right) \quad \middle| \quad s \in \{1, \ldots, n\} \land s \neq \hat{s} \right\}, \tag{1}$$

where $n$ stands for the number of views comprised in the light field. Note that, for the sake of simplicity, the slope $\theta$ is made equivalent to the disparity at the furthest view relative to the reference view. Therefore, in this paper, the terms "slope" and "disparity" are used interchangeably. Moreover, in all our experiments presented here, the central view has been taken for the reference (i.e., $\hat{s} = \lceil n/2 \rceil$). It should also be stressed that the pure linear relationship of Eq. (1) implicitly requires the light-field views to be uniformly sampled within the aperture space and the object's motion to be strictly uniform.

Concerning the construction of the sets $\Omega(u, v, \theta)$, it can be seen that it generally may require an interpolation in the spatial image domain. In our experiments, we have used cubic interpolation in order to achieve results of a higher quality. In cases, where performance is a major concern, one may also employ linear or nearest-neighbor interpolation, which, however, may result in less accurate disparity estimates.

Based on the set $\Omega(u, v, \theta)$, we define the overall cost function $C_{\mathrm{SAD}}(u, v, \theta)$ for each object point $(u, v)$ and any given hypothesized slope $\theta$ as follows:

$$C_{\mathrm{SAD}}(u, v, \theta) = \sum_{P \in \Omega(u,v,\theta)} \left( \sum_{\text{patch pixels}} \left\| P - \mathcal{P}_v^{m\times m}(u, \hat{s}) \right\| \right), \tag{2}$$

where $\| \cdot \|$ usually stands for the $\ell^2$ norm in the case of color and for the absolute difference of radiances for gray-scale data. The inner sum accumulates the pixel-wise differences between two patches of the two views.

As argued before, the cost function $C_{\mathrm{SAD}}$ cannot, by definition, perform reliably in the case of scenes that violate the static Lambertian assumption. Thus, we strive to factor out brightness and contrast variations in different views, which represent the most prominent manifestations of such violations. We do this by means of a pre-normalization of all image patches, which occur in the SAD cost calculation, by subtraction of the mean value of the patch and division by its standard deviation:

$$C_{\mathrm{MSAD}}(u,v,\theta) = \sum_{P\in\Omega(u,v,\theta)} \left( \sum_{\text{patch pixels}} \left\| \mathrm{norm}\,(P) - \mathrm{norm}\left(\mathcal{P}_v^{m\times m}(u,\hat{s})\right) \right\| \right), \tag{3}$$

where $\mathrm{norm}(X) = (X - \mathrm{mean}(X))/\mathrm{std}(X)$ and $\mathrm{mean}(\cdot)$ and $\mathrm{std}(\cdot)$ stand for the mean and standard deviation of radiance values from the given patch, respectively. In the case of color, the normalization is applied to each color channel independently. Otherwise, the MSAD algorithm works exactly the same way as SAD.

Given a slope hypothesis $\theta$, by calculating cost values in all object points $(u, v)$, a cost map is generated, which covers the entire spatial image domain. In order to further stabilize results of block matching and to reduce the influence of noise, we apply spatial filtering to each cost map before identification of the winning hypothesis. Typically, simple box or median filters serve quite well for this purpose. In all our experiments, we have chosen the window size of the box filter equal to the window size used in block matching.

Finally, the winning slope is assessed in each object point $(u, v)$ by accepting the hypothesis, which provides the minimal cost over all considered slopes:

$$\Theta(u,v) = \arg\min_{\theta}\, \widehat{C}(u,v,\theta), \tag{4}$$

where $\widehat{C}(u, v, \theta)$ stands for the spatially filtered cost map $C(u, v, \theta)$. Note that $\theta$ is taken only from a reasonable set of slope hypotheses, which is typically optimized for the given scene.

An important question related to the hypothesis testing is about the choice of a reasonable set of hypotheses to be tested. If there are too few hypotheses, disparity maps cannot be estimated accurately. On the other hand, too many hypotheses result in higher computational demands and, if defined too loosely, they give rise to higher mismatch rates. Therefore, the set of hypotheses has to be chosen carefully to cover the entire expected range of disparities for the given scene, but not more. Another important parameter is the granularity of the hypotheses test set.

In our experiments, we have decided for testing just integer disparities covering the acceptable disparity range. In order to provide finer disparity estimates, we apply a so-called *sub-hypothesis refinement*, which compensates for the lower number of tested hypotheses by means of a quadratic interpolation. In each object point $(u, v)$, we consider the cost estimate $C(u, v, \theta)$ as a function of $\theta$. Then, a quadratic function is fitted locally around the very hypothesis, which provides the minimal cost, and its two neighboring hypotheses from both sides. Finally, the slope, yielding the minimum of the obtained quadratic fit, represents the refined disparity estimate $\widehat{\Theta}(u, v)$. The proposed sub-hypothesis refinement is computationally very cheap, yet efficient compared with an approach where a large number of hypotheses have to be evaluated with a finer disparity step. Moreover, the results show smooth variation of disparities where gentle depth changes appear to be in object space, which only would be achievable with very fine hypothesis test sets, if no sub-hypothesis refinement was considered.

There is one last comment we would like to make on the proposed disparity estimator, in particular about evaluation of confidence of obtained disparity values. In stereo vision, various kinds of confidence measures are applied in order to exclude areas of the scene, where disparity could not be estimated with the required certainty. These measures typically reflect either presence of a useful structure in particular scene locations (as one cannot find correspondences in areas without any structure) or distinction of the best found match from all other tested hypotheses or combination of both (see, e.g, Refs. 13, 15). Of course, such a tool is relevant for our method too, however, we see finding a reliable disparity confidence measure as partly independent from disparity estimation itself. Therefore, we do not pay any special attention to this issue in this paper.

### 3.2 Construction of all-in-focus images

In this section, we describe the construction of all-in-focus images as a method closely related to TDI,[11] a well-known technique implemented in many CCD line-scan image sensors used in machine vision. In order to lower the camera noise, the TDI method integrates over multiple acquisitions of presumably the same object lines collected over consecutive time points. In TDI-enabled image sensors, the integration is implemented in analog circuitry by accumulating electric charges, before the actual A/D conversion takes place. In theory, this allows for increasing the signal-to-noise ratio proportionally with the number of so-called TDI *stages* (i.e., integration steps). Note that TDI is, by definition, (i) only suitable for acquiring planar objects residing entirely in the camera focal plane and (ii) situations where a perfect synchronization between acquisition and transport is guaranteed.

The construction of all-in-focus images follows the same principle as TDI, however, with few important differences. First of all, in our case, the data processing takes place only after all the required data was digitized. This puts strong limitations on the efficiency of noise suppression. However, it enables a locally adaptive integration along optimized EPI slopes, which offers many advantages in return.

Making use of the final disparity estimates from Section 3.1, we define a set of individual radiance values $\widehat{\Omega}(u,v)$ collected from the $v$-th EPI $E_v$ along an imaginary line with the optimized slope $\widehat{\Theta}(u,v)$:

$$\widehat{\Omega}(u,v) = \left\{ E_v \left( u + \frac{s-\hat{s}}{\max(n-\hat{s},\ \hat{s}-1)}\, \widehat{\Theta}(u,v),\ s \right) \ \middle|\ s \in \{1,\dots,n\} \right\}, \tag{5}$$

where $n$ stands for the number of light-field views and $\hat{s}$ is the reference view index. In the next step, the all-in-focus radiance estimate $\Phi(u,v)$ is calculated as the mean of radiance values in $\widehat{\Omega}(u,v)$:

$$\Phi(u,v) = \frac{1}{n} \sum_{r \in \widehat{\Omega}(u,v)} r. \tag{6}$$

In the case of color, the above procedure is applied to each color channel independently, although, in each spatial image coordinate, all the color channels share the same disparity estimate $\widehat{\Theta}(u,v)$.

Let us briefly comment on a theoretical quality of the resulting all-in-focus image. Assuming accurate disparity estimates in all spatial image coordinates, the signal-to-noise ratio of the all-in-focus image is increased approximately by a factor of $\sqrt{n}$ compared with any individual view (i.e., sub-aperture image) comprised in the light field, while preserving the same depth of field. On the other hand, when compared with an image acquired by a standard line-scan camera with some lens $f$-number $N$, the all-in-focus image, obtained by our multi-line-scan method with the $f$-number of $\approx N\sqrt{n}$, would provide the same signal-to-noise ratio as the standard camera. However, the depth of field would be extended approximately by a factor of $\sqrt{n}$. The "square root" rule for combining repeated continuous measurements comes naturally from the central limit theorem, and Witkovsky et. al.[16] showed that this rule also applies to discrete measurements, which is the case for digital imaging. Moreover, apart from defocussing due to the limited depth of field, the TDI-like approaches tend to produce "motion" artifacts in the transport direction when acquired objects depart from the focal plane (e.g., they are of a 3-D nature). Through the locally adaptive integration slopes in EPIs, our all-in-focus images do not suffer from this problem either.

Finally, it should be noted, that the presented approach to all-in-focus image construction is a naive approach, which cannot cope very well with outliers in the integration sets $\widehat{\Omega}(u,v)$. The reason why such a simple method proved to be sufficient for our purposes is, that the outliers are typically caused by occlusions in the scene, which occur very seldom with a narrow-baseline stereo system such as ours.

## 4. RESULTS

In the following, we first provide a comparison of the both considered disparity estimators – SAD and MSAD – based on a synthetic ground truth data. This data violates the static Lambertian assumption and simulates a considerably noisy camera. Afterward, we demonstrate the performance of our experimental multi-line-scan light-field camera, when applied to various real-world objects. A comparison of the all-in-focus images obtained by our method and TDI-like images is provided as well.

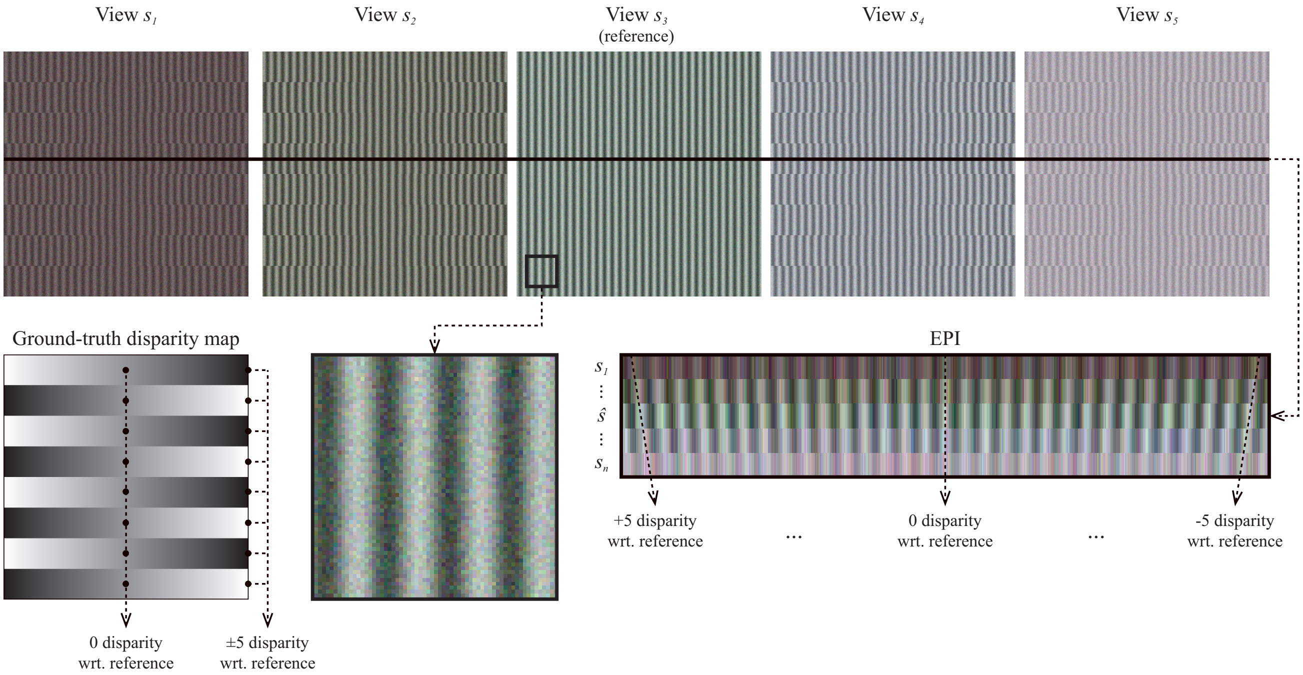

View $s_1$
View $s_2$
View $s_3$
(reference)
View $s_4$
View $s_5$
Ground-truth disparity map
0 disparity wrt. reference
±5 disparity wrt. reference
EPI
$s_1$
$\hat{s}$
$s_n$
+5 disparity wrt. reference
...
0 disparity wrt. reference
...
-5 disparity wrt. reference

Figure 3. Example of the ground truth light-field data comprising 5 views of the signal of 8 times the Nyquist wavelength.

## 4.1 Comparison of SAD and MSAD methods using synthetic data

In this section, we compare the SAD and MSAD methods based on a synthetic data, where the ground truth disparity values are explicitly known. Fig. 3 shows an example of the ground truth light-field data used in this experiment. The performance of both methods was compared on the entire disparity interval of $[-5, 5]$. The light fields consisted of $\{3, 5, 7, 9, 11\}$ views and the signals' wavelengths were $\{2, 4, 8, 16\}$ times the Nyquist wavelength. Block matching and spatial window sizes were equal and chosen from the range $\{3 \times 3, 5 \times 5, 7 \times 7\}$. The camera noise was simulated by an additive Gaussian noise with the amplitude of 10 dB. Properties of a non-static Lambertian scene were mimicked by making the signal contrast and bias considerably different in each view.

Fig. 4 provides a graphical overview of the root mean-squared errors (RMSE) of disparity values delivered by both methods, when compared with corresponding ground truth disparity maps. Tab. 2 lists the results together with a relative comparison of the SAD and our MSAD method expressed in the precision gain of MSAD over SAD (i.e., $(RMSE_{SAD} - RMSE_{MSAD})/RMSE_{SAD}$).

From the graphical overview, it can be observed that the RMSE values of SAD are in general significantly larger (by up to 76%) than the corresponding error values delivered by MSAD (see Fig. 4). A closer look at the values in Tab. 2 reveals that this trend is not always true, in particular for the cases of 2 times the Nyquist wavelength, setups with more than 5 views and domain sizes of $5 \times 5$ and $7 \times 7$, respectively. In these cases, the SAD method shows a slightly better performance than the MSAD method (by up to 6%), which may, however, be considered negligible due to extremely small RMSE values obtained by both methods in all these cases. Therefore, the general dominance of MSAD over SAD is not contested in any of the cases.

Moreover, it is obvious form the graphical overview that the high-frequency signals with wavelengths of 2 times the Nyquist wavelength are handled correctly only when the used light field comprises at least 7 views. For the cases with 3 and 5 views, the obtained disparity precision is very low due to a high chance of a pattern mismatch. On the other hand, when more views are considered, the obtained RMSE values drop dramatically. Therefore, it is advisable to use at least 7 views in our setup.

Finally, the domain size (i.e., the size of image patches used for block matching and spatial filtering) has an influence on the quality of disparity estimation. According to our expectations, larger domain sizes improve the general accuracy of disparity estimates, especially for large wavelengths. It is only natural that the extended supports allow for better capturing those low-frequency structures, which makes block matching more reliable. On the other hand, the larger the block matching domain, the less details are comprised in the resulting disparity maps and the higher computational effort is required. These two aspects have to be taken into account when making decisions about the domain size.

Concerning problematic low-frequency components, that may appear as homogeneous regions w.r.t. the given domain size, such image structures pose a problem for many disparity estimation methods and are usually overcome by means of various pyramid approaches (e.g., Ref. 13) or global optimization techniques (e.g., Ref. 14). Nevertheless, if computation performance is a concern, a pyramid approach is preferable.

## 4.2 Performance of the multi-line-scan light-field camera on real-world objects

In this section, we present disparity maps and all-in-focus images obtained for data acquired by our multi-line-scan light-field camera used to capture real-world objects, such as a 3-D printed staircase, a banknote mapped onto a 3-D printed wave platform, a commemorative coin and a printed circuit board (PCB). All light fields presented in this section comprise 9 views and were acquired with the low-resolution setup except for PCB, which has been acquired with the high-resolution setup (see, Tab. 1, below). Concerning the patch sizes for block matching and spatial filtering, we have employed a smaller domain of 5 × 5 pixels for both the staircase and the coin. For the remaining two objects (i.e., the banknote and PCB), a larger 7 × 7 domain has been used.

### Object 1: 3-D printed staircase

The first presented object is a 3-D printed staircase (see Fig. 5, a), a sketch of the staircase is displayed. In order to provide the block matching with sufficient image structure to work with, we applied a random dot pattern onto the object as well as to its background.

The resulting disparity maps computed with SAD and MSAD are shown in Fig. 6 (a and b), respectively. In both disparity maps, the staircase structure is displayed distinctly, however, the SAD result comprises strong artifacts on the left and right side of the object. These artifacts originate from fast-moving shadows cast by the object during its transport, as the relative position of the object and the illumination is different in each

Table 2. Numerical overview of RMSE values delivered by the SAD and MSAD methods measured against the ground truth model generated for various numbers of views and wavelengths denoted to as $\lambda$. Note that the value of $\lambda$ represents a multiple of the Nyquist wavelength. Configurations providing RMSE accuracy better than 1 are marked gray.

| | | *RMSE of SAD* | | | | *RMSE of MSAD* | | | | *MSAD over SAD precision gain* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Domain size* | $\lambda$ / *Views* | 2 | 4 | 8 | 16 | 2 | 4 | 8 | 16 | 2 | 4 | 8 | 16 | *Avg.* |
| 3 × 3 | 3 | 4.35 | 4.65 | 1.95 | 3.78 | 4.25 | 4.67 | 1.54 | 3.50 | 2% | -1% | 21% | 7% | 8% |
| | 5 | 4.37 | 0.67 | 1.73 | 3.64 | 4.23 | 0.22 | 1.27 | 3.41 | 3% | 67% | 26% | 6% | 26% |
| | 7 | 0.09 | 0.66 | 1.69 | 3.59 | 0.07 | 0.19 | 1.19 | 3.36 | 23% | 71% | 30% | 7% | 33% |
| | 9 | 0.09 | 0.66 | 1.68 | 3.64 | 0.07 | 0.18 | 1.16 | 3.35 | 24% | 73% | 31% | 8% | 34% |
| | 11 | 0.09 | 0.65 | 1.65 | 3.59 | 0.07 | 0.16 | 1.07 | 3.28 | 24% | 76% | 35% | 9% | 36% |
| 5 × 5 | 3 | 4.22 | 4.55 | 1.45 | 3.22 | 4.06 | 4.46 | 0.59 | 2.27 | 4% | 2% | 59% | 30% | 24% |
| | 5 | 4.11 | 0.35 | 1.40 | 3.05 | 3.89 | 0.11 | 0.55 | 2.06 | 5% | 69% | 61% | 33% | 42% |
| | 7 | 0.07 | 0.33 | 1.39 | 3.00 | 0.07 | 0.10 | 0.53 | 1.97 | -2% | 69% | 62% | 34% | 41% |
| | 9 | 0.07 | 0.33 | 1.40 | 3.06 | 0.07 | 0.10 | 0.53 | 1.94 | -2% | 70% | 62% | 37% | 42% |
| | 11 | 0.07 | 0.33 | 1.38 | 3.00 | 0.07 | 0.09 | 0.46 | 1.82 | -2% | 71% | 67% | 39% | 44% |
| 7 × 7 | 3 | 4.06 | 4.42 | 1.21 | 2.90 | 3.82 | 4.25 | 0.35 | 1.44 | 6% | 4% | 71% | 51% | 33% |
| | 5 | 3.81 | 0.14 | 1.18 | 2.80 | 3.51 | 0.10 | 0.40 | 1.32 | 8% | 28% | 66% | 53% | 39% |
| | 7 | 0.07 | 0.14 | 1.18 | 2.76 | 0.07 | 0.10 | 0.40 | 1.25 | -6% | 29% | 66% | 55% | 36% |
| | 9 | 0.07 | 0.14 | 1.18 | 2.83 | 0.07 | 0.10 | 0.41 | 1.25 | -6% | 31% | 65% | 56% | 36% |
| | 11 | 0.07 | 0.13 | 1.17 | 2.77 | 0.07 | 0.09 | 0.36 | 1.12 | -6% | 29% | 70% | 60% | 38% |

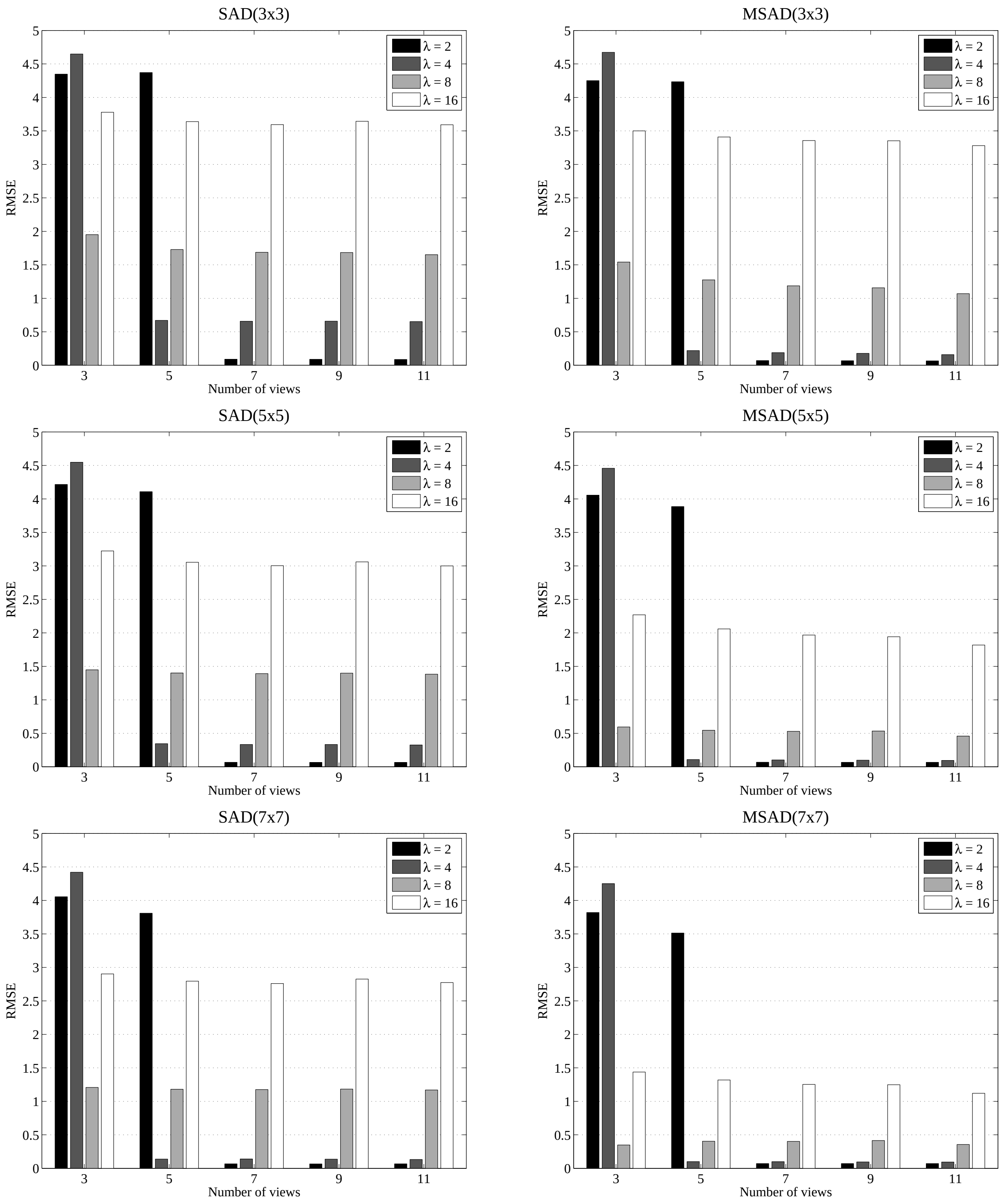


Figure 4. Comparison of the SAD and MSAD methods using synthetic data. The charts show the obtained RMSE values measured against the ground truth model generated for various numbers of views and wavelengths denoted by $\lambda$. Note that the value of $\lambda$ represents a multiple of the Nyquist wavelength. For the actual numerical results shown in this figure, see Tab. 2.

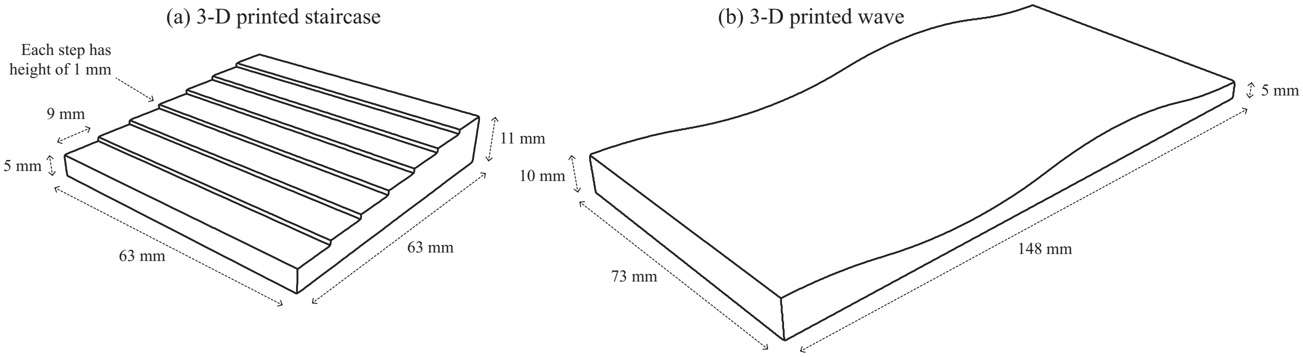


Figure 5. 3-D printed objects: (a) staircase and (b) wave platform.

view. Note that this is a typical example of a non-static Lambertian scene. It is quite natural, that SAD cannot handle these situations correctly, as the algorithm compares original pixel radiance values. Since brightness and contrast are factored out in MSAD, the shadow artifacts do not appear that much in the MSAD disparity map.

Fig. 6 (c) depicts a tentative 3-D visualization of the MSAD disparity map. One can see that the stairs are reproduced very sharply and surfaces are quite smooth with little noise. Indeed, this fine disparity reconstruction was achieved using the low-resolution acquisition setup with the spatial and depth resolution of 120 $\mu$m/px and 2.86 mm/unit disparity, respectively. Note that this is possible only thanks to the employed sub-hypothesis refinement described in Section 3.1. To the right of the staircase, small artifacts can be seen, which originate from the shadow artifacts. Remarkably, one can observe a slight wave structure orthogonal to the staircase slope. This artifact is most likely a result of varying transport speed.

Finally, Fig. 6 (e) shows the all-in-focus image, which has been computed on the basis of the MSAD disparity map. It comprises less noise than the image associated with any single view (see Fig. 6, d) and, moreover, it is considerably sharper than a corresponding image, that would be obtained by a standard TDI approach (see Fig. 6, f) in the areas, where an object goes out of the focal plane.

### Object 2: Banknote mapped onto a 3-D printed wave platform

The second captured object is a banknote that has been applied to a 3-D printed wave platform (see Fig. 5, b). Similarly to the staircase object, the background is comprised of a random dot pattern. The disparity maps computed by SAD and MSAD are shown in Fig. 7 (a and b). Obviously, the SAD result contains substantially more artifacts than the result of MSAD, whereby the main wave structure of the object is well reproduced in both cases. The MSAD method produces artifacts almost exclusively in a homogeneous image region on the left of the banknote (i.e., pink-yellow region in Fig. 7, c and d), while the SAD method has a lot of outliers in regions where there is enough structure for the block matching to operate correctly.

Again, the corresponding all-in-focus image generated by using the disparity map obtained by MSAD (see Fig. 7, c) is essentially sharper than the TDI-like version in Fig. 7 (d).

### Object 3: Commemorative coin

As depicted in Fig. 8, comparable results could be obtained for a commemorative coin. The disparity maps (Fig. 8, a and b) for SAD and MSAD, respectively, reveal that MSAD again produces much less artifacts. Since the coin has a metal surface exhibiting specular reflections at many places, the disparity estimator must be able to handle the gloss and specularity as well. The SAD method often fails in such image regions (e.g., “bright” disparities at the inner coin margin), whereas MSAD tends to suffer from this problem much less. Moreover, MSAD again copes better with the shadow areas to the left and right of the coin.

Concerning the 3-D visualization based on the MSAD disparity map shown in Fig. 8 (c), it again emphasizes the stable quality of the estimate throughout the entire image area.

The reduced noise of the all-in-focus image (see Fig. 8, e) in comparison with a single view image (see Fig. 8, d) is quite visible in this case (see zoomed regions). On the other hand, a difference in sharpness between the all-in-focus image and the TDI-like image, which has been so remarkable in the previous cases, is not distinct in this

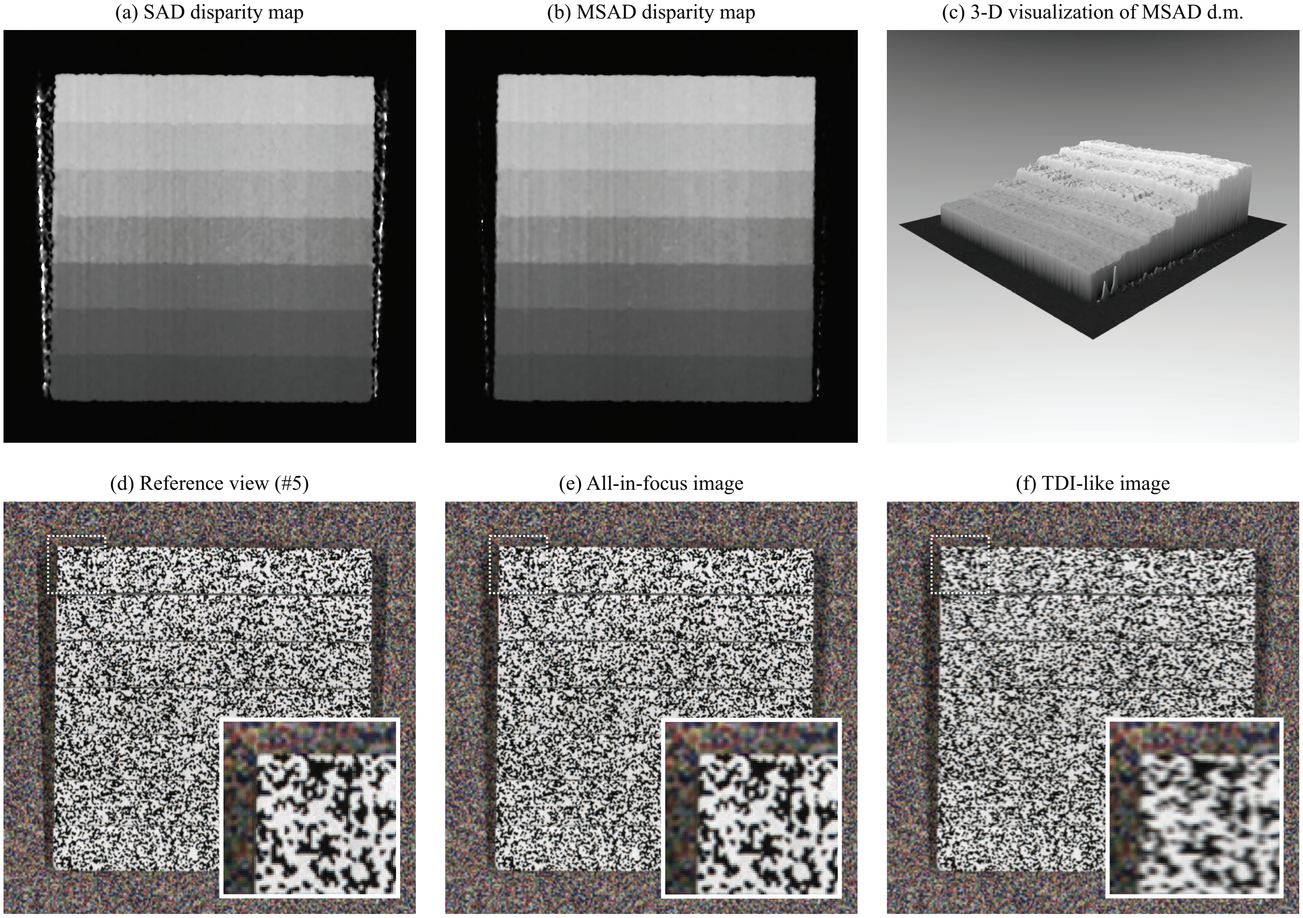


Figure 6. Overview of the results obtained for the 3-D printed staircase.

case anymore (see Fig. 8, e and f), because the 3-D structure of the coin is less pronounced than in the former two examples. Thus, due to smaller deviations from the focal plane, the all-in-focus image and the TDI-like image are more alike. Nevertheless, it should be emphasized that the suggested disparity estimator is capable of recognizing even such fine 3-D structures with such a high precision.

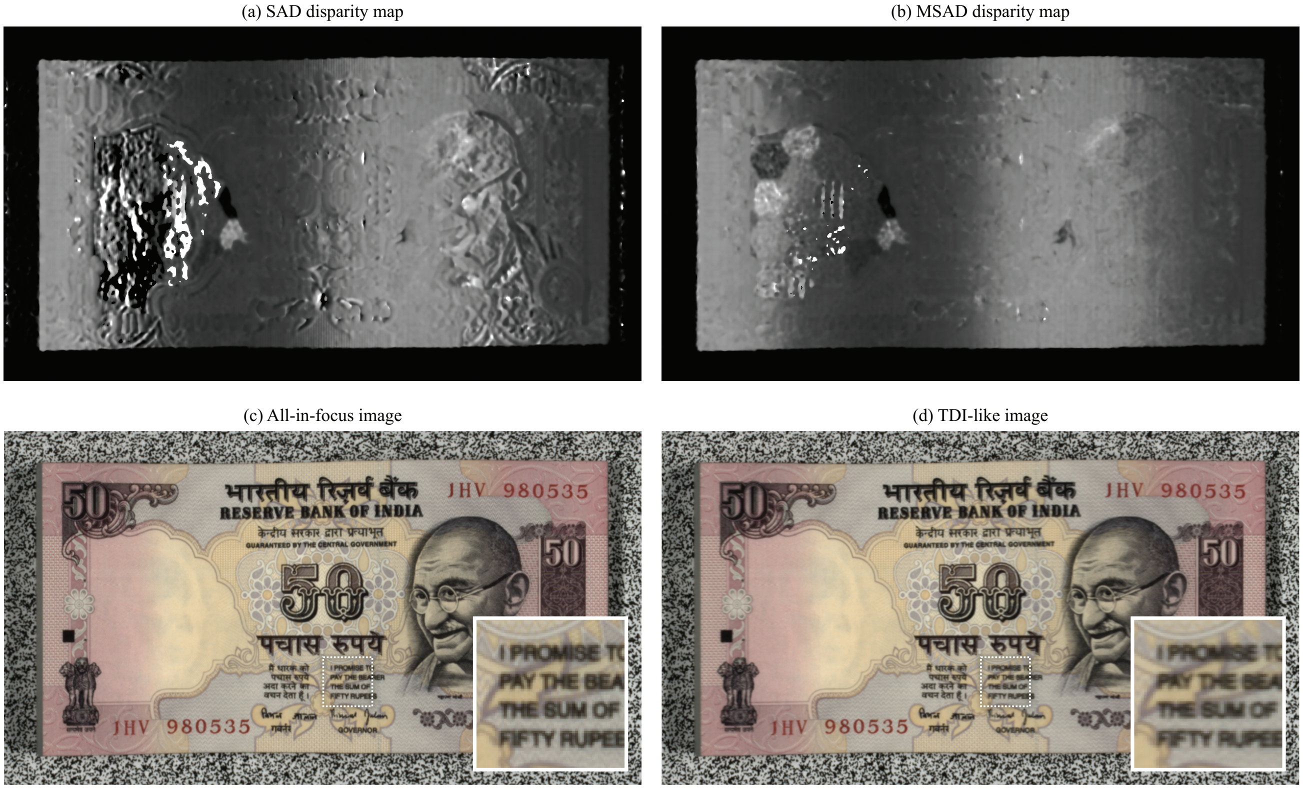
### Object 4: PCB

Finally, we present the results for the printed circuit board of Arduino Due R3 (Fig. 9). Note that in this case we have used the high-resolution configuration of our setup, which provides the spatial and depth resolution of 48 $\mu$m/px and 1.44 mm/unit disparity, respectively.

Once more, the obtained disparity maps (see Fig. 9, a and b) manifest the superiority of MSAD over SAD. The MSAD result contains less outliers and reproduces the PCB 3-D structure with more precision. The 3-D visualization of the MSAD disparity map (see Fig. 9, c) shows how fine structures can still be detected by our method (e.g., little “black” holes over the PCB plate). The reduction of noise in the all-in-focus image w.r.t. a single view image as well as of blur compared w.r.t. the TDI-like image can be clearly observed in this case, too (see Fig. 9, d through f).

## 5. CONCLUSIONS AND DISCUSSION

In this paper, we have presented a multi-line-scan camera setup for capturing a 3-D light field in a line-scan acquisition process. Multiple lines are extracted from a standard area-scan sensor (camera) in each time instance and used to provide different views of an object being transported in front of the camera at a constant speed


(a) SAD disparity map

(b) MSAD disparity map

(c) All-in-focus image

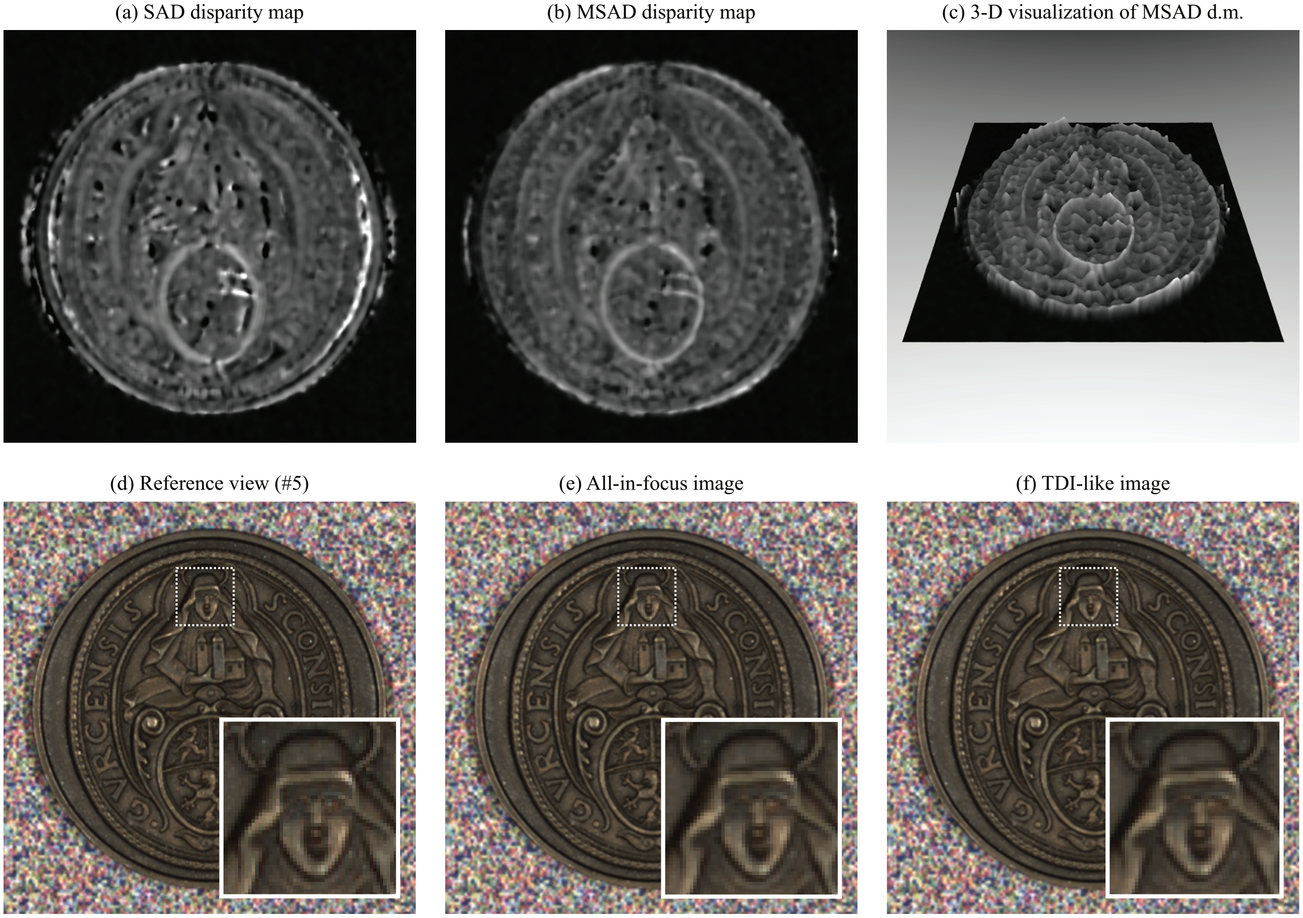
(d) TDI-like image

Figure 7. Overview of the results obtained for the banknote on a 3-D-printed wave platform.

and direction. The proposed acquisition setup is closely related to the TDI approach, however, generalized and implemented entirely in digital instead of analog. Since our system is intended for industrial environments, where stable illumination conditions and pure Lambertian surfaces cannot always be guaranteed, we have proposed a modification of the well known SAD-based block matching disparity estimator – the MSAD approach – which handles objects with glossy and specular surfaces as well as varying illumination conditions better than SAD. The improvement is achieved by factoring out local brightness and contrast variations in the input light field, which are typical for scenes violating the static Lambertian assumption. Afterward, we described an approach to the construction of all-in-focus images, which improve the signal-to-noise ratio and reduces the motion blur produced by a standard TDI approach when objects depart from the focal plane of the camera.

We have demonstrated the superiority of MSAD over SAD in couple of experiments with synthetic light-field data, where the ground truth disparity values were explicitly known. In the synthetic light fields, the camera noise and the lightness variations have been modeled to mimic scene conditions that violate the static Lambertian assumption. Moreover, we have shown that with our concept of the multi-line-scan light-field camera combined with the MSAD method it is possible to achieve impressively precise disparity maps not only for the synthetic data, but also for a set of real-world objects. Finally, the obtained disparity estimates have been used for generating all-in-focus images which have significantly reduced the amount of TDI motion artifacts, while the signal-to-noise ratio has been improved considerably with respect to any individual view.

Last but not least, it should be mentioned that another important advantage of our setup is its form factor as well as price when compared with conventional camera arrays. As the proposed multi-line-scan light-field camera is essentially just a single fast area-scan camera, the final system is maximally rigid w.r.t. the alignment of multiple cameras in the array. Just as there is no free lunch, the crucial point of our system becomes the synchronization between the acquisition and the transport. If the synchronization fails, the obtained disparity maps would blend the actual 3-D structure of the acquired object with transport artifacts, which may obstruct an accurate 3-D reconstruction. Nevertheless, this problem has almost no impact on the all-in-focus image construction, which will still be generated correctly.



(a) SAD disparity map

(b) MSAD disparity map

(c) 3-D visualization of MSAD d.m.

(d) Reference view (#5)

(e) All-in-focus image

(f) TDI-like image

Figure 8. Overview of the results obtained for the commemorative coin.

## ACKNOWLEDGMENTS

This work has been partially supported by the Slovak Grant Agency for Science (project No. 2/0043/13).

## REFERENCES

[1] Bolles, R. C., Baker, H. H., David, and Marimont, H., "Epipolarplane image analysis: an approach to determining structure from motion," *International Journal of Computer Vision* **1**(1), 7–55 (1987).

[2] Raskar, R., Tumblin, J., Mpohan, A., Agrawal, A., and Li, Y., "State of the art report (STAR): computational photography," in [*Proc. of ACM/Eurographics*], (September 2006).

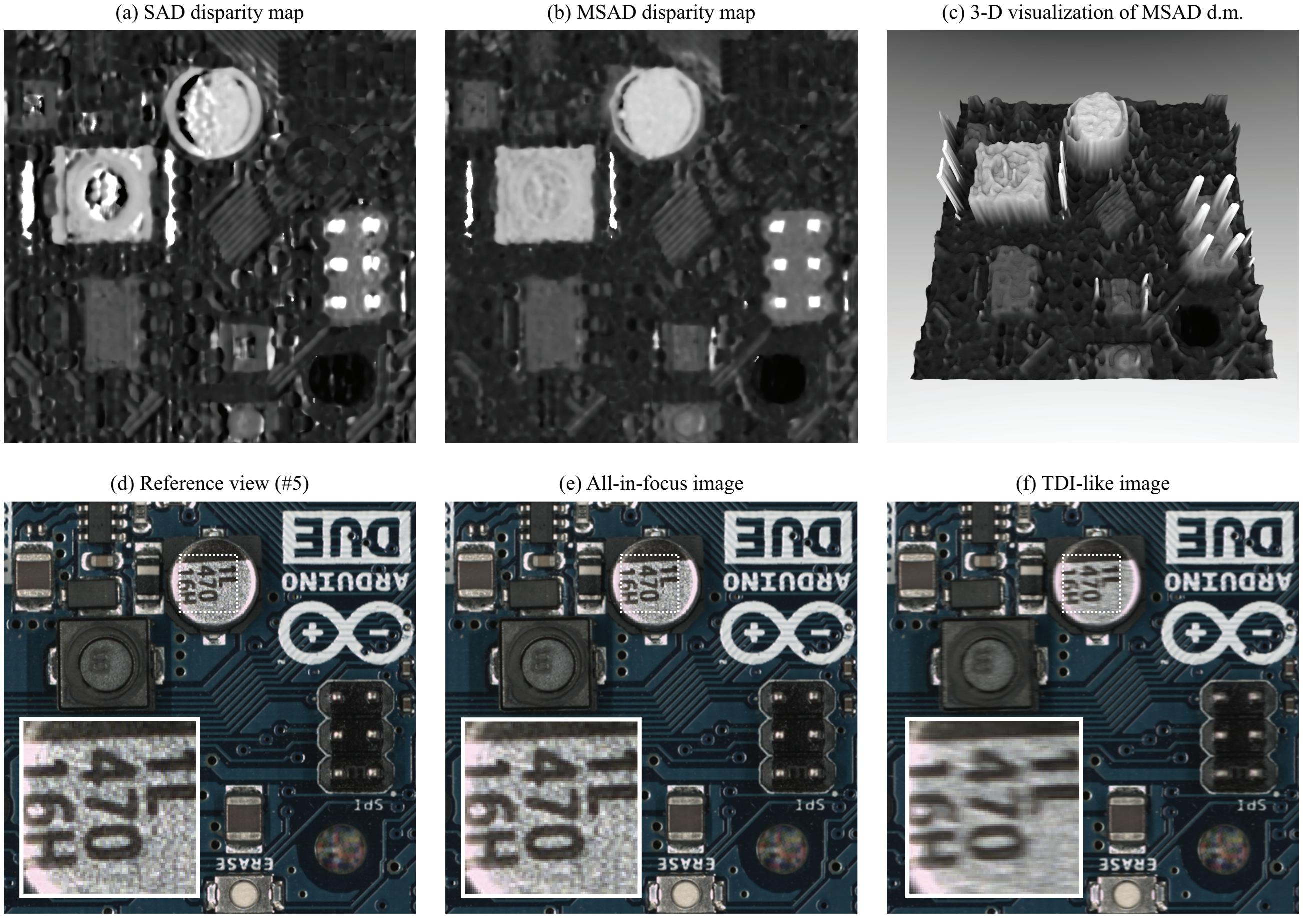
[3] Levoy, M. and Hanrahan, P., "Light field rendering," in [*Proc. of Conference on Computer Graphics and Interactive Techniques*], *SIGGRAPH*, 31–42 (1996).

[4] Wilburn, B., Joshi, N., Vaish, V., Talvala, E.-V., Antunez, E., Barth, A., Adams, A., Horowitz, M., and Levoy, M., "High performance imaging using large camera arrays," *ACM Trans. Graph.* **24**, 765–776 (July 2005).

[5] Venkataraman, K., Lelescu, D., Duparré, J., McMahon, A., Molina, G., Chatterjee, P., Mullis, R., and Nayar, S., "PiCam: an ultra-thin high performance monolithic camera array," *ACM Trans. Graph.* **32**(5) (2013).

[6] Davis, A., Levoy, M., and Durand, F., "Unstructured light fields," *Computer Graphics Forum* **31**, 305–314 (May 2012).

[7] Liang, C.-K., Lin, T.-H., Wong, B.-Y., Liu, C., and Chen, H., "Programmable aperture photography: multiplexed light field acquisition," *ACM Transactions on Graphics* **27**(3), 55:1–55:10 (2008).


(a) SAD disparity map

(b) MSAD disparity map

(c) 3-D visualization of MSAD d.m.

(d) Reference view (#5)

(e) All-in-focus image

(f) TDI-like image

Figure 9. Overview of the results obtained for the printed circuit board of Arduino Due R3.

[8] Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., and Hanrahan, P., "Light field photography with a hand-held plenoptic camera," Tech. Rep. CSTR 2005-02, Stanford University (April 2005).

[9] Perwaß, C. and Wietzke, L., "Single lens 3D-camera with extended depth-of-field," in [*Proc. of SPIE: Human Vision and Electronic Imaging XVII*], **8291**, 829108–829108–15 (2012).

[10] Holländer, B., Štolc, S., and Huber-Mörk, R., "Multi-view line-scan inspection system using planar mirrors," in [*Proc. of SPIE Optical Metrology: Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection*], **8791**, 879118–879118–11 (2013).

[11] Wong, H.-S., Yao, Y., and Schlig, E. S., "TDI charge-coupled devices: design and applications," *IBM Journal of Research and Development* **36**, 83–105 (January 1992).

[12] Goldlücke, B. and Wanner, S., "The variational structure of disparity and regularization of 4D light fields," in [*Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 1003–1010 (2013).

[13] Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., and Gross, M., "Scene reconstruction from high spatio-angular resolution light fields," *ACM Transactions on Graphics* **32**(4), 73:1–73:12 (2013).

[14] Wanner, S. and Goldlücke, B., "Globally consistent depth labeling of 4D light fields," in [*Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 41–48 (2012).

[15] Ambrosch, K., Zinner, C., and Kubinger, W., "Algorithmic considerations for real-time stereo vision applications," in [*Proc. of Machine Vision and Applications (MVA)*], 231–234 (May 2009).

[16] Witkovský, V. and Wimmer, G., "Interval estimation of the mean of a normal distribution based on quantized observations," *Mathematica Slovaca* **59**(5), 627–645 (2009).