Light Field from Smartphone-based Dual Video

Bernd Krolla[1], Maximilian Diebold[2], Didier Stricker[1]

[1]German Research Center for Artificial Intelligence, Kaiserslautern, Germany
[2]Heidelberg Collaboratory for Image Processing, Heidelberg, Germany

Abstract. In this work, we introduce a light field acquisition approach for standard smartphones. The smartphone is manually translated along a horizontal rail, while recording synchronized video with front and rear camera. The front camera captures a control pattern, mounted parallel to the direction of translation to determine the smartphones current position. This information is used during a postprocessing step to identify an equally spaced subset of recorded frames from the rear camera, which captures the actual scene. From this data we assemble a light field representation of the scene. For subsequent disparity estimation, we apply a structure tensor approach in the epipolar plane images.
We evaluate our method by comparing the light fields resulting from manual translation of the smartphone against those recorded with a constantly moving translation stage.

1 Introduction

While processing capabilities and hardware specifications of todays smartphones approach those of classical desktop computers, they are additionally equipped with a wide set of various sensors.

Besides multiple processing units and high amounts of memory, the latest smartphones are typically provided with GPS, IMUs, compass, (stereo) cameras, and other sensors.

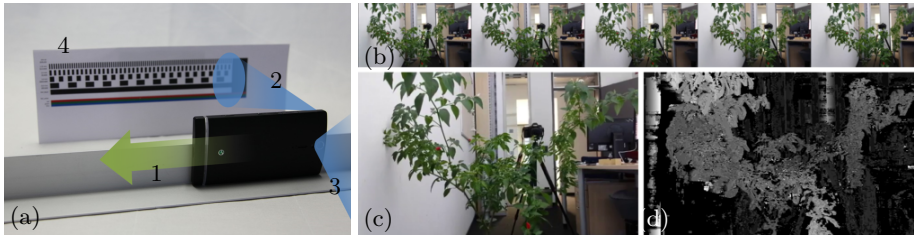

Fig. 1: (a) We manually translate a smartphone along a horizontal rail (1) while recording synchronized video with front (2) and rear camera (3). The front camera captures a control pattern (4) allowing the identification of the smartphones position while the main camera captures the actual light field data (b) of the scene (c). The resulting depth-map is shown in (d).

Currently ongoing research within the domain of depth sensing technologies [6, 10] including new camera systems such as [5, 8, 9, 22] will most likely introduce an additional set of sensors for smartphones in the near future.

Besides this research and development of future devices, most of todays produced smartphones are typically equipped with at least one rear camera at the backside, as well as a front camera, which faces towards the user. Setting up on this hardware configuration, we aim to perform light field acquisition in an easy and end-user friendly manner.

We therefore introduce a new acquisition approach for light fields exploiting the availability of front and rear cameras of todays smartphones (Figure 1). In this context, different methods to precisely localize the smartphone during the light field acquisition are evaluated and discussed.

2 Related work

A light field can be represented by the plenoptic function, introduced by Adelson and Bergen [1], Levoy and Hanrahan [12] and McMillan *et al.* [15].
The plenoptic function gives the fundamental understanding of representing and acquiring light fields e.g as 2D light field representation called Lumigraph, introduced by Gortler *et al.* [7]. Since then different methods have been established to exploit all information a light field provides.
Veeraraghavan *et al.* [21] introduce a light field acquisition camera using aperture masks. This masks attenuates the incident light rays without refracting them. Purpose of these masks is the modulation of the captured images. The light field is achieve by applying a Fourier transform based image demodulation.
An alternative approach also using aperture masks is called programmable aperture. Lian *et al.* [13] applying mask based multiplexing exploiting the fast multiple-exposures of cameras to generate the light field datasets.
In contrast to digital approaches, Levoy and Hanrahan [12] acquire light field data using a single moving camera. This is the simplest method and utilizes a computer controlled 3D translation stage called Gantry to capture suitable images for light field image processing.
A very similar approach is structure from motion introduced by Bolles *et al.* [2], having a straight-line camera motion system to capture a dense sequence of images. In their paper, they also introduce the exploitation of the epipolar plane images to obtain information about the three-dimensional position of objects and its usability.
Aside single moving cameras also large camera arrays as introduced by Wilburn *et al.* [25] are a possibility to capture light field datasets. While camera arrays for light field acquisition mostly have the constraint to be mounted on a planar grid with equidistant spacings between the cameras, Snavely *et al.* [19] introduces a method to reconstruct 3D object having a unstructured collection of images of the same object. The introduced system automatically computed the view point of each camera and generates a sparse 3D model of the scene and image, while the images can be captured in a random way.

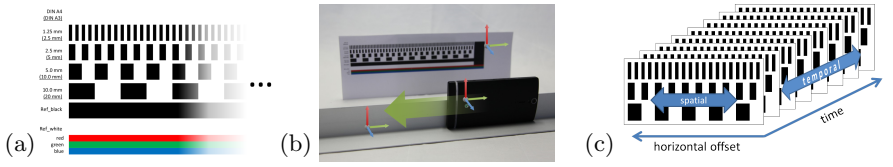

Fig. 2: (a) shows a part of the repetitive control pattern, which was used to locate the smartphones position during the horizontal translation (b). The captured video stream of this camera results in a set of video frames (c), which was used to localize the smartphones position

Similar to that Davis *et al.* [3] present a system to interactively acquire light fields using a hand-held commodity camera. The system has real-time feedback to the photographer to obtain a dense light field of the captured object for the 3D reconstruction. An other possibility to capture wide angle light fields is been introduced by Taguchi *et al.* [20]. In this paper a spherical catadioptric cameras is modeled, using mirror balls mounted on a common plane. For the capturing, an aligned camera to the mirror set is been used to obtain the light field data. While the above introduced methods are based on customary perspective cameras Adelson and Bergen [1] and Ng *et al.* [16, 17] introduce so called plenoptic cameras having a micro lens array in front of the image sensor to obtain beside spacial information also angular information of the scene. Unfortunately, the obtained angular information is always combined with a reduction of spacial resolution. Thus Lumsdaine and Georgiev [4, 14] and Perwass *et al.* [18] introduce focused plenoptic cameras. Difference to the already introduced plenoptic cameras is the changed focus position of the main lens. Thus a higher resolution in the resulting light field images is obtained, but also the computational effort is much higher.

The work of Levoy [11] provides a smartphone application, which allows the generation of computational images with a narrow depth of field. While the application is characterized by its good usability, a generation of disparity maps of the scene is not performed.

3 Method

Assuming to be provided with a smartphone, the required hardware setup was chosen to allow for a low-cost and end-user friendly light field acquisition. The presented approach is ready to be used with any *state-of-the-art* smartphone, which is able to capture *dual video* with its main (=*rear*) and sub (=*front*) camera as shown in Figure 1(a).

In this work, we used a smartphone, which records synchronized dual video with 24*fps* and a resolution of 640 × 360 pixel per video stream. Neglecting the additional setup for evaluation as detailed in Section 4, further necessary equipment is limited to a rail allowing for horizontal sliding of the phone, as well as a control pattern provided as a simple printout on a paper sheet.

The control pattern as pictured in Figure 2(a) is horizontally subdivided into multiple binary patterns of different frequencies. This periodicity allows for an easy determination of the relative camera position, while excluding any absolute positioning of the camera towards the pattern. We achieved with the given layout an easy and fast processing leading to sufficient positioning results.

The actual capturing of the light field is shortly demonstrated in the supplementary video material and consists in the manual shift of the smartphone along the horizontally mounted rail.

The recording was done at a relatively low translation velocity ($\leq 3cm/s$) to allow for a dense sampling of the scene through the video frames and to avoid a degradation of the recorded data through motion blur or influence of the rolling shutter.

3.1 Key frame extraction

Being provided with the dual video stream of synchronized front and rear camera, we now aim to describe the captured light field information with a sparse subset of video frames to make it available for subsequent light field processing. To do so, we need to identify a set of equally spaced frames within the main video sequence.

Having the simultaneously recorded video stream of the front camera at hand, a wide variety of approaches is applicable to perform this task, which consists in the analysis of a two-dimensional space with a spatial and a temporal dimension as indicated in Figure 2(c).

In this work, we confine ourselves to evaluate the following methods:

Spacial Intensity Change around a fixed key position (SIC) When applying this approach, each frame of the front video stream is considered independently to retrieve information about the smartphones relative position towards the control pattern. The intensity gradient in the direction of translation is thereby computed at a preselected position as shown in Figure 3(a).

As soon as the calculated gradient exceeds a given threshold τ_g, indicating the passage of an intensity border within the binary control pattern, the corresponding video frame of the rear-camera is added to the light field representation. The value of τ_g was hereby identified as one third of the difference between the reference values for black and white intensity.

$$\tau_g = \frac{1}{3} \cdot \frac{i_{white} - i_{black}}{2} + i_{black}. \tag{1}$$

Being provided with those preselected keypoints, this approach allows an online detection of relevant frames while the video-capturing is still in progress.

Temporal Intensity Change on fixed key position (TIC) For this approach, we choose keypoints within the video stream of the front camera in the same manner as for the SIC approach (See Figure 3). However, the introduced

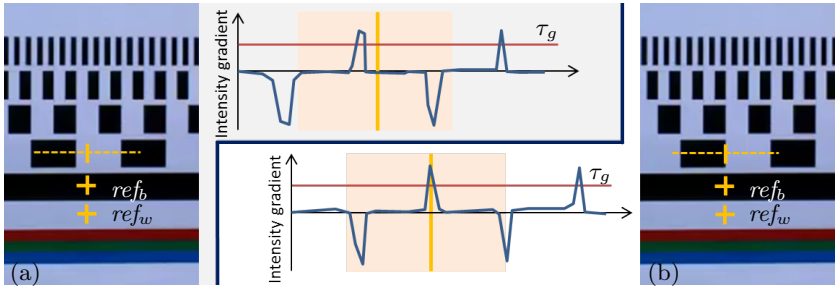

Fig. 3: (a) and (b) show two exemplary frames of the control pattern, captured by the front camera. A keypoint with accompanying evaluation window is indicated by the topmost orange mark. The two marks below are used to extract corresponding intensity values for white (ref_w) and black color (ref_b). Gradients of the intensity values for the two frames are assessed within the evaluation window (center) leading to a keyframe detection for frame (b).

evaluation window as not used, but the intensity values at those keypoint positions were extracted for all frames of the video. To identify equidistant frames for light field parametrization, we then detected the edges of the binary control pattern at the chosen keypoints by comparing intensity values between current and proceeding frame. Whenever the difference of those intensities exceeded a given threshold τ the corresponding frame of the rear-camera was marked to be part of the light field.

The threshold τ was hereby computed in two different ways within this work: Assuming the overall intensity maximum i_{max} and minimum i_{min} to be given, we calculated a static threshold τ_s as average in a straight-forward manner by

$$\tau_s = \frac{(i_{max} + i_{min})}{2}. \tag{2}$$

While this threshold is easily determined during a postprocessing step, a temporally dynamic threshold τ_d is obtained by smoothing the intensity distribution with a Gaussian function. For frame i we obtain through discrete convolution of the intensity function f_{int} with the Gaussian distribution g:

$$\tau_d(i) = (f_{int} * g)(i) = \sum_m f_{int}[m]g[i-m] \tag{3}$$

Equidistant frames in time-domain To allow for a comparison of the presented methods for key frame extraction (SIC and TIC), we applied a further approach for keyframe extraction which is independent from the recorded video stream of the front camera (control pattern).

We used this method exclusively in conjunction with the translation stage and extracted a subset of frames by identifying every nth frame of the rear camera, while assuming constant translation velocity.

3.2 Light field processing

The captured keyframes of the main camera were rectified using the calibration approach of Vogiatzis *et al.* [23] and can be represented as three dimensional light field volume, utilizing the two plane parametrization as introduced by Gortler [7] called lumigraph. Thus we define the Π-plane containing the focal points $s \in \Pi$ of all cameras and the Ω-plane which denotes the image coordinates $(x, y) \in \Omega$. The resulting three dimensional light field volume becomes

$$L : \Omega \times \Pi \to \mathbb{R} \qquad (s, x, y) \mapsto L(s, x, y), \tag{4}$$

where $L(s, x, y)$ defines the color in each point.

In the resulting light field data is an epipolar plane image obtained by slicing through the light field volume. To achieve this the parameter y is set to a constant value $y*$. The resulting epipolar plane image is then defined by the function

$$S_{y^*} : \Sigma_{y^*} \to \mathbb{R} \tag{5}$$

$$(x, s) \mapsto S_{y^*}(x, s) := L(s, x, y^*). \tag{6}$$

An epipolar plane image contains information about the scene depth in terms of depth dependent orientations, see Figure 4.

To analyze these orientations, we use the structure tensor

$$J = \xi * \begin{pmatrix} \left(\frac{\partial \hat{S}}{\partial x}\right)^2 & \frac{\partial \hat{S}}{\partial x} \cdot \frac{\partial \hat{S}}{\partial s} \\ \frac{\partial \hat{S}}{\partial s} \cdot \frac{\partial \hat{S}}{\partial x} & \left(\frac{\partial \hat{S}}{\partial s}\right)^2 \end{pmatrix} =: \begin{pmatrix} J_{xx} & J_{xs} \\ J_{xs} & J_{ss} \end{pmatrix} \tag{7}$$

with the abbreviation

$$\hat{S} := \sigma * S_{y^*}, \tag{8}$$

where σ and ξ define a Gaussian smoothing. The resulting scene disparity information can now be computed as given in [24] using the equation

$$d = \tan\left(\frac{1}{2} \arctan\left(\frac{2J_{xs}}{J_{xx} - J_{ss}}\right)\right), \tag{9}$$

where only the structure tensor components are used to compute the underlying orientations.

Fig. 4: Example of an Epipolar Plane Image (EPI) assembled from 31 images

Fig. 5: (a) On site capturing setup: A tripod-mounted and battery-powered translation stage allows for horizontal camera shifts with welldefined velocities as well as for manual operation.

4 Evaluation and Results

To evaluate the proposed approach we exploit the dual capturing mode of a *state-of-the-art* smartphone for parallel video acquisition of front and rear camera.

Besides a manual translation of the smartphone along a rail (Figure 1(a)), also a translation stage as shown in Figure 5 was used to capture light field data of different scenes. This stage provides a translation range of more than 25cm and operates highly accurate regarding the precision of velocity and positioning.

While the shifting velocity during manual operation could not be measured precisely, the velocity of the translation stage was set to a constant value of 7 mm/s during the acquisition process.

We evaluated besides the office scene (Figure 1) two outdoor scenes, while applying manual and automatic translation techniques. Figures 6 and 7 provide an overview of the obtained results, while the performance of the introduced keyframe detection approaches is discussed below.

Spacial Intensity Change around a fixed key position (SIC) Since all frames in this approach are evaluated independently from each other, it cannot be avoided that directly consecutive frames are detected as keyframes for the light field: While the gradient detection implies the evaluation of the keypoints neighborhood, it occurs, that consecutive frames are selected as key frames (e.g. for the 2.5mm pattern in Figure 7(c)), especially at low translation speeds of the camera and high recording frequencies.

Temporal Intensity Change on fixed key position (TIC) This approach uses the two previously introduced thresholds τ_s and τ_d. For large parts of the evaluated scenarios, both approaches deliver very similar results (Figures 6 and 7)

Since the acquisition of the light fields requires constant illumination conditions, the property of the dynamically calculated threshold τ_d to adopt to

possible illumination changes within the scene is obsolete. Additionally tends this technique to deliver unreliable results at the start and end points of the recorded video streams. For the generation of disparity maps, resulting from the TIC approach, we therefore used exclusively the simpler static thresholding technique basing on τ_s.

Equidistant frames in time-domain This approach for keyframe extraction relies on constant translation velocity of the smartphone and was applied for evaluation purposes exclusively in conjunction with the translation stage. Since this method does not exploit any information from the control pattern, we used it to assess the results of the TIC and SIC approch (Figure 6 and 7).

5 Discussion

While providing in this work a conceptual overview over the proposed light field acquisition approach, we observed a variety of aspects, which currently prevent further improvement of results.

The recording with two independently managed (front and rear) camera systems complicates a full parameter control. Both cameras were checked to capture frames synchronously, while further camera parameters such as focus, white-balance or ISO-values remain uncorrelated. Establishing a strongly coupled camera pair, which assures the named parameters to be mutually controlled would allow to exploit especially prepared control pattern for global white-balancing.

During the evaluating of manually captured scenes, we furthermore noticed a high sensitivity of the light field processing methods against camera shakes, which require the user for careful acquisition. Image registration techniques as part of the postprocessing could possibly reduce this demand.

6 Conclusion

In this work, we introduced a light field acquisition approach for standard smartphones exploiting synchronized dual video capturing of front and rear camera. We evaluated the proposed method for different scenes and achieved comparable results for the proposed TIC keyframe extraction approach and the equidistant frame extraction method, relying on the capturings with the translation stage.

7 Acknowledgements

This work was funded by Sony Deutschland, Stuttgart Technology Center, EuTEC and is a result of the research cooperation between STC EuTEC, the Heidelberg Collaboratory for Image Processing (HCI) and the German Research Center for Artificial Intelligence (DFKI).

We would like to thank Thimo Emmerich, Yalcin Incesu and Oliver Erdler from STC EuTEC for their feedback to this work and all the fruitful discussions.

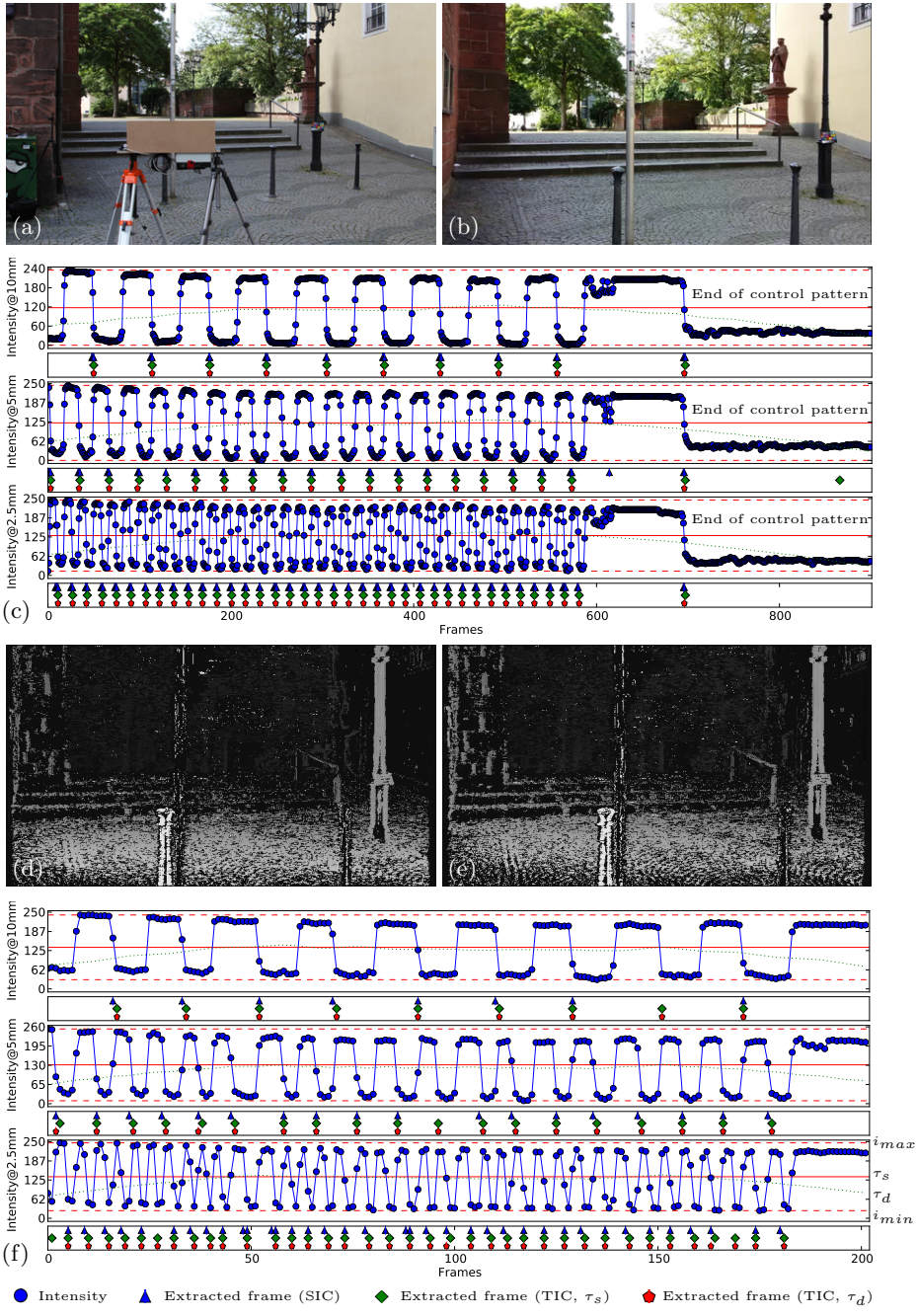

Fig. 6: (a) On site acquisition setup. (b) Exemplary frame captured by the smartphones rear camera. (c) Temporal intensity distribution for different keypositions in the front camera stream with an overview of extracted keyframes for different extraction methods (automatic translation). Resulting disparity maps for automatic translation (d), using the equidistant extraction approach and for the TIC extraction approach (e), using the $5mm$ control pattern. (f) Extracted keyframes using manual translation.

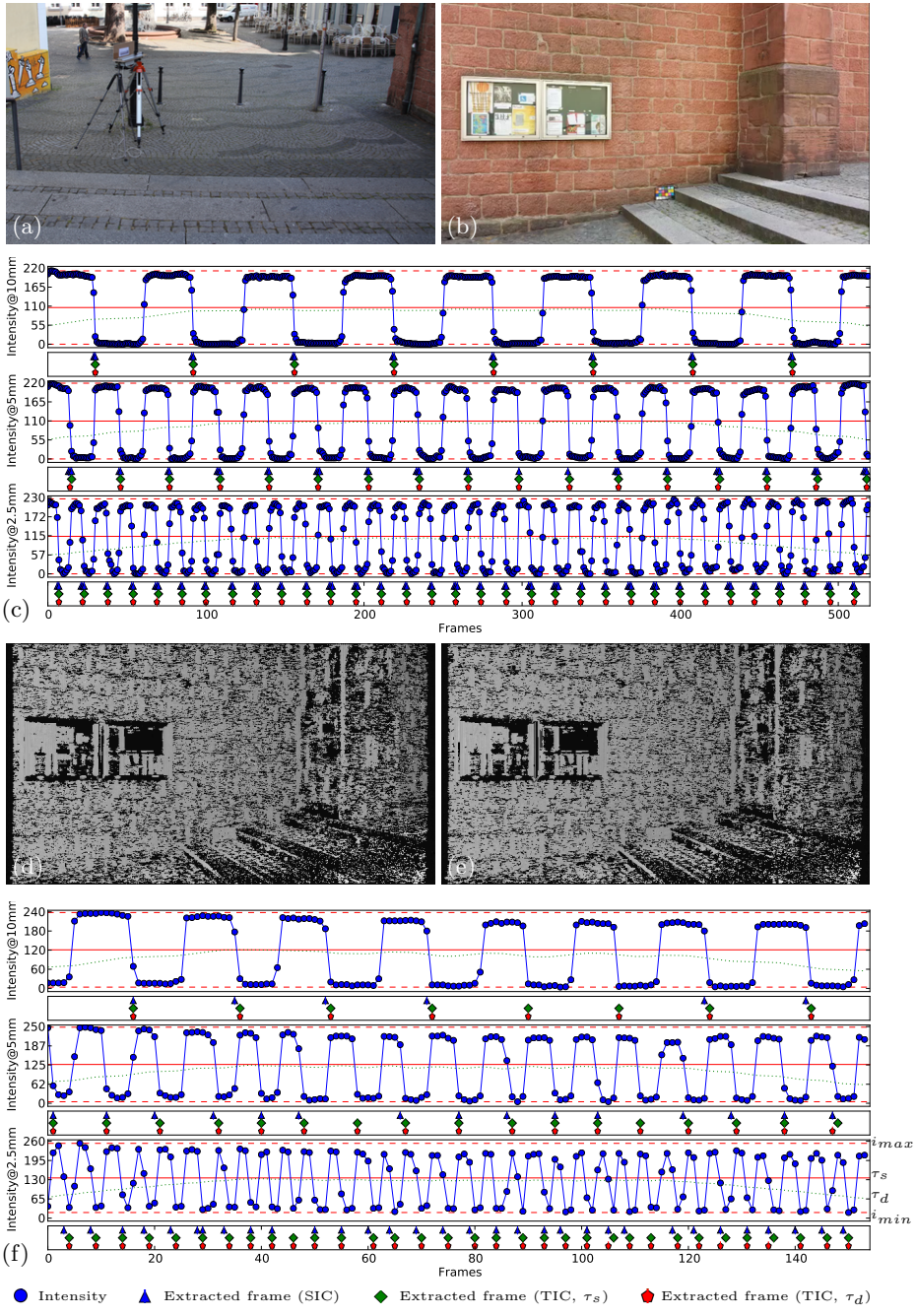

Fig. 7: (a) On site acquisition setup. (b) Exemplary frame captured by the smartphones rear camera. (c) Temporal intensity distribution for different keypositions in the front camera stream with an overview of extracted keyframes for different extraction methods (automatic translation). Resulting disparity maps for automatic translation (d), using the equidistant extraction approach and for the TIC extraction approach (e), using the 2.5mm control pattern. (f) Extracted keyframes using manual translation.

References

1. E.H. Adelson and J.R. Bergen. The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1:43–54, 1991.
2. R.C. Bolles, H.H. Baker, and D.H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.
3. A. Davis, M. Levoy, and F. Durand. Unstructured Light Fields. *Comp. Graph. Forum*, 31(May 2012):305–314, 2012.
4. T. Georgiev and A. Lumsdaine. Focused plenoptic camera and rendering. *Journal of Electronic Imaging*, 19:021106, 2010.
5. Raytrix GmbH. Raytrix, 2014. `http://www.raytrix.de/`.
6. Google. Project tango, 2014. `https://www.google.com/atap/projecttango/`.
7. S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen. The Lumigraph. In *Siggraph*, 1996.
8. HTC. Htc one m8, 2014. `http://www.htc.com/us/smartphones/htc-one-m8/`.
9. Lytro Inc. Lytro, 2014. `https://store.lytro.com/`.
10. Occipital Inc. Structure sensor, 2014. `http://structure.io/`.
11. M. Levoy. Synthcam, 2014. `https://sites.google.com/site/marclevoy/`.
12. M. Levoy and P. Hanrahan. Light field rendering. pages 31–42, 1996.
13. C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. Chen. Programmable Aperture Photography: Multiplexed Light Field Acquisition. 27(3):1–10, 2008.
14. A. Lumsdaine and T. Georgiev. The Focused Plenoptic Camera. In *In Proc. IEEE Int. Conference on Computational Photography*, pages 1–8, 2009.
15. L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. pages 39–46, 1995.
16. R. Ng. *Digital Light Field Photography*. PhD thesis, Stanford University, 2006. Note: thesis led to commercial light field camera, see also `www.lytro.com`.
17. R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. Technical Report CSTR 2005-02, Stanford University, 2005.
18. C. Perwass and L. Wietzke. The Next Generation of Photography, 2010. `www.raytrix.de`.
19. N. Snavely, S. Seitz, and R. Szeliski. Photo Tourism: Exploring image collections in 3D. 2006. `http://phototour.cs.washington.edu/bundler/`.
20. Y. Taguchi, A. Agrawal, S. Ramalingam, and A. Veeraraghavan. Axial Light Fields for Curved Mirrors: Reflect your Perspective, Widen your View. 2010.
21. A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin. Dappled Photography: Mask Enhanced Cameras for Heterodyned Light Fields and Coded Aperture Refocussing. 26(3):1–69, 2007.
22. K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar. Picam: an ultra-thin high performance monolithic camera array. *ACM Transactions on Graphics (TOG)*, 32(6):166, 2013. `http://www.pelicanimaging.com/technology/`.
23. G. Vogiatzis and C. Hernández. Video-based, real-time multi-view stereo. *Image and Vision Computing*, 29(7):434–441, 2011.
24. S. Wanner and B. Goldluecke. Variational Light Field Analysis for Disparity Estimation and Super-Resolution. 2013.
25. B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Transactions on Graphics*, 24:765–776, July 2005.