Heterogeneous Dataset Acquisition for a Continuously Expandable Benchmark (CEB)

Bernd Krolla
DFKI - German Research
Center for Artificial Intelligence
bernd.krolla@dfki.de

Didier Stricker
DFKI - German Research
Center for Artificial Intelligence
didier.stricker@dfki.de

ABSTRACT

Ongoing research within the field of computer vision yielded a wide range of image based 3D reconstruction approaches. Starting years ago with low resolution RGB images as input, we face today a wide and fast growing range of available imaging devices to perform this task.
To allow for a good comparability of resulting reconstructions, many different benchmarks and datasets have been made available. At the same time, we observe, that these benchmarks commonly address only a single capturing approach omitting the chance to compare against results of other acquisition methods.
In contrast to such homogeneous benchmarks, we present in this work a heterogeneous benchmark, considering different acquisition devices to obtain our datasets. Besides these datasets, we furthermore provide reference data for download.
To lastly keep track of the rapidly increasing number of different acquisition sensors, we opt to provide occasional updates of this benchmark within the future.

Keywords

Computer Vision, 3D Reconstruction, Benchmark, Heterogeneous Dataset Acquisition

Within the field of computer vision, image-based 3D reconstruction of objects and environments has been subject to intense research since many years. Gradually, the estimation of essential and fundamental matrices [7], camera calibration [20] and multiple view reconstruction [7] was understood and improved [11, 12, 16, 17].
Having calibrated camera parameter as well as sparse pointclouds of a scene at hand, many different reconstruction algorithms have been developed, to generate notable image based reconstruction results such as [2, 3, 5, 15].
While most of the former approaches for image-based 3D reconstruction rely on the processing of perspective RGB-images, the computer vision community can nowadays access a rapidly expanding variety of new sensors:
Recent developments introduced technical devices such as *high definition* and *4K* video-cameras, *high dynamic range (HDR)* imaging devices, *RGB-depth (RGBD)* cameras, consumer cameras capturing at frame rates of *90Hz* and more, *stereo* cameras, *light field* cameras, *Time of Flight (ToF)* cameras and many more. Various of those devices are capable to offer new approaches for 3D reconstruction, which are commonly addressed in the context of ongoing research. To access and quantify the potential of such newly developed algorithms, a wide range of benchmarks has been made available [4, 6, 8, 9, 14, 18, 19].
When taking these above listed benchmarks for 3D reconstruction into consideration, we claim that they do not yet allow for a comparison of reconstruction algorithms, which rely on different acquisition approaches: A benchmark for RGB-image based 3D reconstruction allows for a comparison of different algorithms which rely on RGB-data. But the very same benchmark excludes any performance assessment with respect to approaches which apply RGB-D, lightfield or video data. This lack of comparability of reconstruction approaches is therefore the underlying motivation for the publication of our presented benchmark.

Contribution We introduce in this work a benchmark consisting of datasets captured from a small set of objects by applying a heterogeneous variety of acquisition sensors. We furthermore aim at a continuous expansion of the dataset by making acquired data from new devices available in the future.
The core contribution within this work is therefore summarized as follows: We selected a set of objects, which provide different challenges for 3D reconstruction We

captured datasets using imaging devices of different kind and quality and make those publicly available. We provide reference data using a structured light approach [1].

Nomenclature Within the further course of this work, all acquisition devices used for the actual capturing are referred as *devices*.
Any physical subject, which has been acquired by a *device* is referred as *object*. It is noteworthy, that an object can therefore also be assembled from multiple, jointly mounted items.
A *dataset* furthermore refers to the digitalized data resulting from an acquisition process of an *object* by using a *device*.
An *environment* is finally considered in this work to contain all surroundings of the acquisition setup, such as background or illumination situation.
All datasets are finally combined into one benchmark, presented in this work. Since the authors intent to add future datasets, whenever new acquisition devices become available, the benchmark may be referred as *CEB (Continuously Expandable Benchmark*).

1 RELATED WORK

In [18], Seitz provided the well-known Middlebury benchmark, consisting of a main set of 2 different objects, acquired from 317 different camera positions while supplying according camera parameter. Reconstruction results obtained from the provided images can be submitted and benchmarked against ground truth data.
Furukawa provided a similar scenario as benchmark in [4]. While providing images with a significantly higher resolution and calibrated camera parameter, ground truth is not provided for all datasets.
Jensen introduced a benchmark in [8], which provides in contrast to the aforementioned benchmarks a wider range of acquired objects, while using a 6-axis industrial robot for the image acquisition. Moreels provided in [14] a benchmark, containing 3D objects on a turntable under varying illumination conditions. The dataset however, remains without ground truth data.

Ground truth vs. Reference Within a complete and meaningful benchmark, the careful generation of an accurate ground truth is however always an important point. While benchmarks, which rely on synthetically generated data are capable of providing ideal ground truth data, any measurement based ground truth acquisition is always subject to error prone measurements.
The different implications in this context are discussed in detail by Kondermann in [10].
To minimize the occurring errors to a minimal ratio, Seitz [18] combined more than 200 laser scans of a single object and applied super resolution algorithms for an improved overall result. At the same time, the actual images of the dataset were provided at a relatively low resolution of 640x480 pixel.
Strecha provided in [19] a laser scan as ground truth for their reconstruction challenges. In their work, they estimated the expected precision of the acquired scans and supplied the ground truth together with an estimated variance of the obtained 3D points.

2 OUTLINE

The remainder of this paper is organized as follows: In Section 3, we detail the choice of objects, which were used within the benchmark. Subsequently we elucidate the acquisition process of the individual datasets in Section 4 and present a set of reference measurements. We discuss and conclude this work in Section 5. For further material the reference may be made to the supplementary material, submitted in conjunction with this work. The benchmark itself is available at `http://ceb.dfki.uni-kl.de`.

3 DATASET COMPOSITION

The presented benchmark consists of a total set of 11 different objects, containing various reconstruction scenarios and challenges.
Parts of these objects are composed from groups of items, other datasets consist of single objects.
An important prerequisite to all selected objects is the expected longterm usability for reconstruction purposes to comply with the previously introduced option for future expansion of the *CEB* by adding further datasets.
To satisfy this requirement, exclusively rigid objects were selected to be part of the benchmark. The different objects themselves unify furthermore various geometric challenges including repetitive structures, self occlusions, smooth, irregular, convex and concave surfaces. The benchmark is furthermore characterized by various different surfaces subsumed by the different objects, including wood, plaster, painted plaster, plastics, metal, Styrofoam and others.
In summary, Table 2 provides an overview over the different objects and their main characteristics, while Table 1 gives an overview over the naming conventions for the accompanying camera parameter.

4 ACQUISITION PROCESS

4.1 Preparation

Preceding to the first data acquisition, all objects were mounted on top of quadratic base plates with an edge length varying between 10*cm* and 30*cm*, acknowledging the varying overall size of the objects as listed in Table 2.
Each plate contains a set of drilled holes to allow for a precise mounting on different underground and environments. To assure a stress-free mounting, the objects

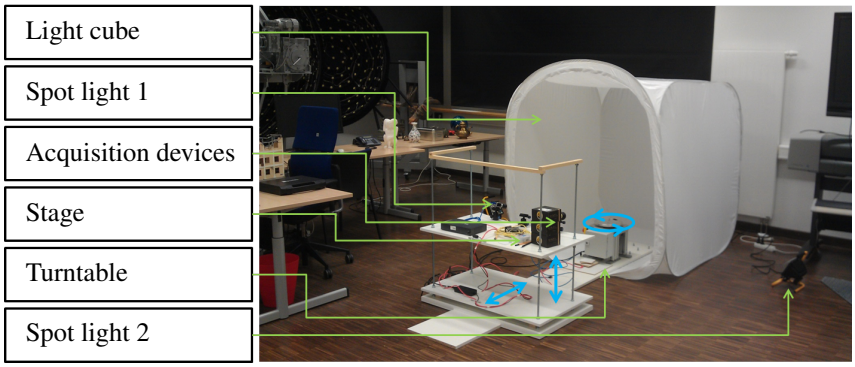

Figure 1: Main acquisition environment: A turntable mounted inside a light tent, being surrounded and illuminated by a set of point light sources. The whole setup is software controlled, placed within a windowless room and allows for a positioning of objects with a precision of 0.012°.

were not skewed onto the mounting plates. Instead, all objects were fixed using a low temperature glue, which avoided thermal dependent tensions during the cooling process.
Taking these measures into account, the authors consider the objects to be ready to meet the requirements for long term availability.

4.2 Dataset acquisition

The wide majority of the provided datasets was acquired within an indoor environment with constant and controlled acquisition conditions. We aimed hereby towards a good reproducibility of external parameters and comparability between different camera types in respect to various acquisition characteristics, such as point of view, number of acquired images and illumination conditions.
To assure the comparability of view points onto the objects for different camera types, all mounting plates with their attached objects were setup on top of a turntable as depicted in Figure 1. The turntables intrinsic positioning precision allowed hereby to approach 240'000 different equally distributed positions in the course of a single 360° turn leading to an angular resolution of 0.0015° with a positioning uncertainty of 0.012° as stated by the manufacturer. This positioning mode was used to line up the objects and to acquire images with the varying imaging devices.
The turntables rotation mode was applied to capture videos with varying devices. The objects rotation was hereby captured at varying velocities in the range of 12'000, 10'000, 8'000, 6'000, 4'000, 2'000 motor steps per minute (corresponding to $\frac{1}{3}$, $\frac{2}{5}$, $\frac{1}{2}$, $\frac{2}{3}$, 1 and 2 *rpm*).

Camera calibration Preceding to each dataset acquisition, a calibration of camera parameter was conducted as proposed by Vogiatzis and Hernández in [21]. The resulting images with the calibration pattern are provided along with the retrieved intrinsic parameter provided for the download.

Illumination To ensure a well defined and reproducible illumination situation, we chose a windowless room for the object acquisition to be independent from any daylight changes. The turntable with the mounted objects was placed inside a light tent, which served as light diffuser. The illumination of the setup was then provided by 3 point light sources. The choice of halogen lamps allowed for a natural illumination compared to narrowband LED-spectra.

Illumination documentation To allow for a color calibration of the individual capturing devices, we acquired

Table 1: Exemplary listing of provided extrinsic and intrinsic parameter for a DSLR-camera. Note: *fx,fy,cx,cy* and α are provided as a joint camera matrix **K**, together with a distortion matrix **D**. Other camera types, such as lightfield cameras, depth or stereo cameras are provided with their individually adapted setting.

Name	Type	Description
hd	Extrinsic	Horizontal distance between cameras principle point and turn table base
vd	Extrinsic	Vertical distance between cameras principle point and turn table base
fx	Intrinsic	Cameras focal length, expressed in pixels
fy	Intrinsic	Cameras focal length, expressed in pixels
cx	Intrinsic	Horizontal coordinate of the cameras principle point
cy	Intrinsic	Vertical coordinate of the cameras principle point
α	Intrinsic	Skew value of the camera sensor

 ISBN 978-80-86943-65-7

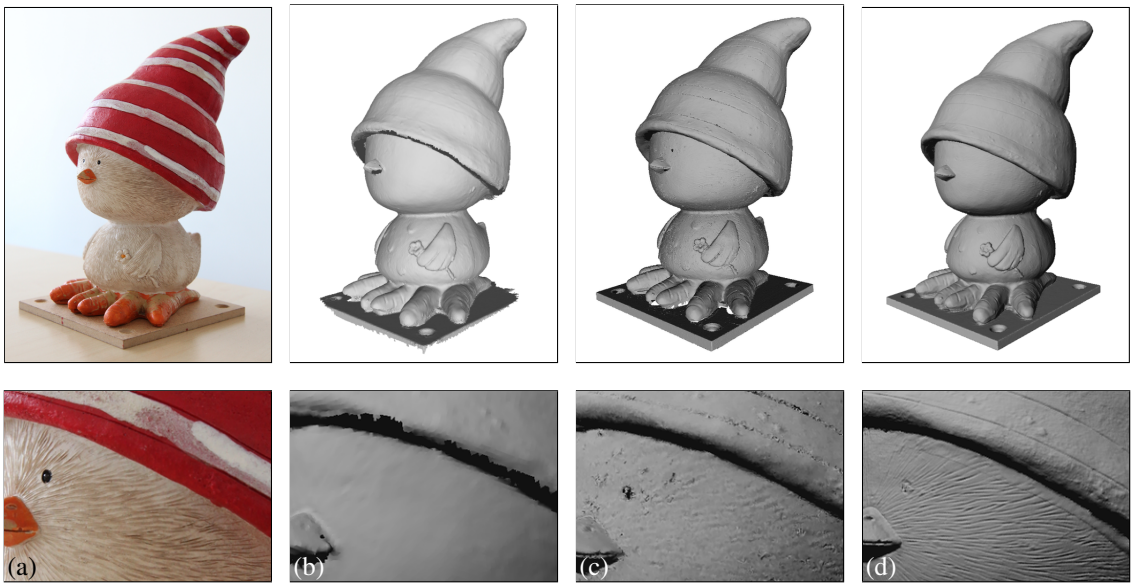

Figure 2: Detail of different acquisition approaches for reference generation: Closeup of the dataset (a). Resulting mesh from the hand held artec scanner acquisition consisting of 620k faces (b). Resulting mesh of the laser scanning approach consisting of 2.05M faces (c). Resulting mesh from the structured light approach as provided by [1] consisting of 45.9M faces (d).

a small set of images of a Macbeth ColorChecker board [22] before capturing the actual datasets. These images allow for a color calibration of the devices, being able to compensate automatic white balancing as enabled by some of the capturing devices. These acquired images are made available within the dataset, without using them to correct for any color balancing of the dataset images.

Logging and documentation To allow for a good understanding of the utilized camera setup and a good traceability of the performed steps and actions during the dataset acquisition, we setup and used a set of logging tools to check and log various types of data and parameters. Using this approach, we assured the documentation of each dataset acquisition with respect to currently chosen type of scene illumination, selected camera parameters and further information.

Further environments Some of the provided datasets were acquired in different environments: To add further characteristics to the benchmark, a small subset of datasets was acquired in a non reproducible manner with limited control of the environmental conditions. Exemplary, we refer to the provided freehand acquisitions in an outdoor environment, which expands the variety of reconstruction scenarios, but depends heavily on the experimenters camera handling and the current weather conditions, making it practically impossible to exactly reproduce an identical scenario for further capturings with different cameras.

4.3 Acquisition devices

The overall set of employed acquisition devices sums up to 7 different devices, while some of those were used for the acquisition of multiple datasets, differing in terms of acquisition mode and acquisition environment: One might consider DSLR cameras used in an *indoor* acquisition scenario in *video* mode and in *outside* acquisition scenarios taking hand held *images* of a dataset.
In general, the acquisition devices can be split up into different groups taking different characteristics into consideration:

Active vs. passive The majority of the applied acquisition devices is characterized by its passive acquisition process, exploiting exclusively incoming illumination emitted by the scene itself.
Active acquisition devices, characterized by their emission of sampling patterns are commonly susceptible to strong surrounding illumination. We therefore did not acquire any outdoor datasets using active acquisition devices. For the standardized indoor acquisition process, however, we used the probably most prominent representative, Microsofts Kinect 360 [13], which relies on the emission of a dot-pattern within the infrared frequency domain.

Image vs. video capturing The presented benchmark provides image-based as well as video-based datasets.
The image-based dataset acquisition of different objects consists hereby in a number of 200 images

 ISBN 978-80-86943-65-7

Figure 3: Visualization of the current download interface: The different datasets are organized in an array with respect to the different objects and the acquisition device (left). Datasets can be selected for download in three different ways: Column-based to download all datasets of a certain acquisition device, row-based to download all datasets of a certain object and individually selected to download a specific device-object-combination (right).

for the main acquisition environment, acquired at well defined object positions.
Some of the devices were ready to allow for video as well as image capturing. For those, we provide image and video based datasets.
Information, referring to the images or videos, such as resolution, frame rate or compression algorithms is provided with the individual datasets.

Mounting Regarding the mounting of the acquisition devices, we intended to satisfy two different demands: To allow for a good reproducibility, most acquisition devices were rigidly connected with a solid acquisition stage as shown in Figure 1. To handle acquisition scenarios, as performed by end users, we furthermore added video and image datasets with hand held acquisition.

For several acquisition devices, such as light-field cameras, special considerations were respected to provide an appropriate acquisition scenario. These considerations are then listed within the corresponding logging files of the datasets.
A complete tabular overview of all devices, which were considered for the dataset acquisition is provided in Table 3. For each dataset, we specify therein a set of extrinsic and intrinsic parameter, which is downloadable along with the imaging data.

4.4 Reference acquisition

In order to allow for a meaningful evaluation of different reconstruction approaches, we provide 3D models of the objects. To acquire those, we considered a variety of different approaches:

Artec Spider The reference data, which was acquired with this hand held 3D scanner was our first approach to provide 3D models of the objects. The resulting models were generated from multiple partial scans, which were aligned against each other using provided software.
Manual operation however results in SLAM-like acquisition approach, while the translation of the device during the acquisition leads possibly to a less precise registration of the camera positions (See Figure 2(b)).

Industry scanner We provided the objects furthermore to a laser scanning supplier, leading to reconstruction results as shown in Figure 2(c).

Structured light scaning approach Best reconstruction accuracies however were achieved using a structured light approach [1] as shown in Figure 2(d).

Figure 2 provides an overview over the provided reference datasets. Visual inspection demonstrates the varying level of reconstructed details for the different approaches.
Complying with the previously stated concept to provide an *Continuously Expandable Benchmark*, we do not consider these reconstructions as ground truth (implying to provide *perfect* data, but aim to provide reference reconstructions (*as good as possible*), leaving room to possible future improvements of reconstruction algorithms.

5 RESULTS AND DISCUSSION

We make the benchmark, as a result of the previously detailed acquisition work publicly available to the

 ISBN 978-80-86943-65-7

computer vision community. Acquired datasets as well as the introduced reference data is provided for download as shown in Figure 3.
We furthermore aim to occasionally provide new datasets to the benchmark within the future.

6 ACKNOWLEDGEMENTS

The authors would like to thank Norbert Schmitz for the turntable setup, Johannes Köhler for the reference generation, Bertram Taetz for the support in context of the Raytrix capturing, Moshin Munir for the Kinect and outdoor acquisitions and Santosh Shah for his implementation of the web interface.
The work was carried out during a research cooperation between the Computational Imaging Group at the Stuttgart Technology Centre of Sony Deutschland GmbH and the German Research Center for Artificial Intelligence (DFKI). We would like to thank in particular Yalcin Incesu and Oliver Erdler from Sony Stuttgart for their feedback and fruitful discussions. This work was cofunded by the BMBF-project DENSITY (01IW12001).

REFERENCES

[1] 3digify.com. http://www.3digify.com, 2015.

[2] Christian Bailer, Manuel Finckh, and Hendrik PA Lensch. Scale robust multi view stereo. In *Computer Vision–ECCV 2012*, pages 398–411. Springer, 2012.

[3] Simon Fuhrmann and Michael Goesele. Floating scale surface reconstruction. *ACM Transactions on Graphics (TOG)*, 33(4):46, 2014.

[4] Y. Furukawa and J. Ponce. 3d photography dataset http://www-cvr.ai.uiuc.edu/ponce_grp/data/mview/, May 2006.

[5] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1362–1376, 2010.

[6] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

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

[8] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanaes. Large scale multi-view stereopsis evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 406–413, June 2014.

[9] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus H Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.*, 32(4):73, 2013.

[10] Daniel Kondermann. Ground truth generation http://resources.mpi-inf.mpg.de/conferences/up2013/up2013_files/up2013-abstracts/kondermann/daniel-kondermann.pdf, 2013.

[11] Quan-Tuan Luong and Olivier D Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17(1):43–75, 1996.

[12] Paulo RS Mendonça and Roberto Cipolla. A simple technique for self-calibration. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 1999.

[13] Microsoft. Microsoft Kinect 360 http://www.xbox.com/en-US/kinect.

[14] P. Moreels and P. Alatorre. 3d objects on turntable http://www.vision.caltech.edu/pmoreels/Datasets/TurntableObjects/.

[15] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3248–3255, Dec 2013.

[16] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999.

[17] Fabio Remondino and Sabry El Hakim. Image based 3d modelling: A review. *The Photogrammetric Record*, 21(115):269–291, 2006.

[18] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 519–528. IEEE, 2006.

[19] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. 2008.

[20] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. *Vision algorithms: theory and practice*, pages 153–177, 2000.

[21] George Vogiatzis and Carlos Hernández. Automatic camera pose estimation from dot pattern, http://george-vogiatzis.org/calib/, 2010.

[22] xRite Inc. Colorchecker classic http://xritephoto.com/, 2014.

Table 2: Table of the acquired objects

Name	Picture	Plate, Height	Material	Geometry	Reflectivity	Further characteristics
Buddha		10x10cm, $\approx 8cm$	Coated plaster	Minor occlusions	Highly reflective	Due to the high reflectivity, the capturing environment tends to impact the reconstruction
Chicken		10x10cm, $\approx 18cm$	Plastics	Moderate complexity	Diffuse	Feather-like surface contains chamfers $< 1mm$ width, self-occlusions under the had brim
Cup		10x10cm, $\approx 9cm$	Glazed ceramic	Smooth and even	Specular surface in untextured regions	Untextured, specular concave interior of the cup, minor self-occlusions due to cups handle
Dragon		15x15cm, $\approx 20cm$	Coated plaster	Complex microscopic structures of the surface.	Diffuse	Surface contains chamfers $< 1mm$ width, contains partial self-occlusions
Elephant		15x15cm, $\approx 13cm$	Coated plaster	Moderate complexity	Highly reflective	Surface contains chamfers $< 1mm$
Elk		10x10cm, $\approx 15cm$	Wood	Moderate complexity	Diffuse	Antlers introduce self-occlusions
Mbdhct		30x30cm, $\approx 10cm$	(Painted) wood, plastics, ceramics, metal	Usage of multiple objects causes self-occlusions	Different types, mostly diffuse	Nomenclature: Mole, Box, Duck, Home (sweet Home), Clock, Teapot
Metal-objects		15x15cm, $\approx 16cm$	Metal (also plastic and Styrofoam®)	The usage of multiple objects causes self-occlusions	Mainly metallic surface implying reflectivity	The implied screw thread represents a highly repetitive pattern.
Owl		15x15cm, $\approx 25cm$	Thin metal sheets, implies minor self-occlusions	Painted metal sheets, transparent glass eyes	Diffuse. Exception: eyes	Upper body is flexibly mounted onto the lower body, allowing for nonrigid dataset acquisition
Santa		10x10cm, $\approx 15cm$	Painted clay	Smooth surface without occlusions.	Diffuse	Feature based reconstruction approaches might work best for high resolution images, which resolve minor texture variations of the object
Scw		30x30cm, $\approx 19cm$	Usage of multiple objects causes self-occlusions	Contains partially transparent surfaces	Metallic, transparent and semi-transparent surfaces	Contains repetitive, structures (Threads). Nomenclature: Shampoo, (CPU)-cooler, Wifi-card.

 ISBN 978-80-86943-65-7

Table 3: Table of considered acquisition devices (open to future extensions).

Device	Picture	Full name	Specification	Characteristics
Bloggie		Sony Bloggie 3D	Full HD Stereo camera	Provides full-hd stereo images (1920x1080, mpo, jpg) and hd video (mp4, 1920x1080px @30fps)
Eos5		Canon EOS5 Mark II	Professional DSLR camera	Provides raw (cr2) and jpg images (resolution 5616x3744pixel)
Eos500		Canon EOS500	DSLR camera	Provides raw (cr2) and jpg images (resolution 4752x3168pixel) and full-hd video (mov)
HTC		HTC Desire HD	Smartphone	Provides hd video (3gp, 1280x720px @30fps)
Kinect		Microsoft Kinect Xbox 360	RGBD camera	Provides frames of 640x480px @30fps (when capturing video)
Raytrix		Raytrix R5	Light field camera	4.2 Megarays, 2048x2048pixel @25fps (GigE)
Techsolo		Techsolo TCA-4810 Webcam	Webcam	640x480 @15fps (avi)