The City of Sights: Design, Construction, and Measurement of an Augmented Reality Stage Set

Lukas Gruber [1]* Steffen Gauglitz [2] Jonathan Ventura [2] Stefanie Zollmann [1] Manuel Huber [3]
Michael Schlegel [3] Gudrun Klinker [3] Dieter Schmalstieg [1] Tobias Höllerer [2]

[1]Graz University of Technology [2]University of California, Santa Barbara [3]Technische Universität München

Figure 1: Left, center, right: The original Arc de Triomphe, a virtual model and a miniature paper model. Middle left and middle right: Various views of the City of Sights, showing a virtual and a real representation of the total assembly.

Abstract

We describe the design and implementation of a physical and virtual model of an imaginary urban scene — the "City of Sights" — that can serve as a backdrop or "stage" for a variety of Augmented Reality (AR) research. We argue that the AR research community would benefit from such a standard model dataset which can be used for evaluation of such AR topics as tracking systems, modeling, spatial AR, rendering tests, collaborative AR and user interface design. By openly sharing the digital blueprints and assembly instructions for our models, we allow the proposed set to be physically replicable by anyone and permit customization and experimental changes to the stage design which enable comprehensive exploration of algorithms and methods. Furthermore we provide an accompanying rich dataset consisting of video sequences under varying conditions with ground truth camera pose. We employed three different ground truth acquisition methods to support a broad range of use cases. The goal of our design is to enable and improve the replicability and evaluation of future augmented reality research.

1 Introduction

Many Augmented Reality (AR) applications require a model or "stage" for a testbed, user interface, or demo purposes (cf. Figure 2 for a few examples). Unfortunately, each researcher has to go through the process of finding an appropriate model him- or herself, and especially for comparisons, it would be desirable to have the same model.

While certain datasets and models have been established as defacto standards and widely used in certain domains in the computer vision and computer graphics communities (e.g., [1, 23, 32]), this has not yet happened for AR, although the need for common datasets and quantitative evaluations has been recognized [10, 21, 33].

There might be one important practical reason for this: while the above datasets consist only of either image or video data or a digital 3D model, AR research often requires a physical model (see Figure 2), and ideally both a digital and physical model. We are not aware of any freely available model that would fulfill these requirements.

In this paper, we will present our approach to provide such a dataset — called the "City of Sights" — and discuss the design choices that were involved. We begin in Section 2 by reviewing existing datasets which have been used in AR and related areas. In Section 3, we discuss research areas within AR that would benefit from a physical+virtual model, and derive a set of general requirements for design and creation of such a model from these. We arrive at a simple and cheap solution for creating a physical representation of the data set by using paper models, as described in Section 4. To provide easily available measurements of the entire data set we recorded several video sequences with ground truth. We discuss this data and our ground truth acquisition methods in Section 5. We summarize our dataset in Section 6 and discuss conclusions in Section 7.

2 Related Work

Computer vision: image/video sequences. There exist many datasets and evaluation frameworks for different computer vision problems, such as Quam's Yosemite sequence [4] and the Middlebury dataset [1] to evaluate optical flow methods, and Seitz et al. [32]'s dataset for multi-view reconstruction.

Several image datasets have been used for evaluating interest point detectors and local invariant descriptors [6, 22, 23, 29]. Here, the data consists of planar images subjected to different geometric and photometric transformations such as rotation, scale, noise, JPEG compression, brightness variations and others. Moreels and Perona [24] used images of 100 physical models seen from 144 different viewpoints to conduct a similar evaluation on complex 3D objects.

To evaluate visual tracking, evaluation frameworks and datasets of video sequences depicting planar tracking targets have been used [10, 21, 39]. Targets are chosen to exhibit different levels of texture richness and self-similarity (repetitive textures). Similar image- or video-based datasets have been used for tracking-by-detection [20] and optimization of natural feature targets [11].

Computer graphics: digital 3D models. The two probably most famous and widely used computer graphics models are the

*e-mail: lgruber@icg.tugraz.at

Figure 2: AR models/"stages" used by (from top left to bottom right) Lepetit and Berger [19], Schöning et al. [30], Pan et al. [26], Raskar et al. [28], Bandyopadhyay et al. [2]. All images courtesy of the respective authors.

"Teapot", initially created by Newell [25] in 1974/75, and the "Stanford Bunny", created by Turk and Levoy [36] in 1994 (although its creators did not grant it a picture in its debut paper). The Teapot, which was created manually and consists of only a few faces, now serves as demo object in virtually every OpenGL demo and became an icon for the SIGGRAPH community [34]. The Stanford Bunny features a detailed surface structure and is thus popular as test object especially for mesh creation algorithms. Further objects may be found for example in the Stanford 3D scanning repository[1].

Augmented reality: physical models. As AR draws from many different research fields including computer vision and graphics, many of the above mentioned datasets and resources are clearly valuable to different AR applications. However, even if the data stems from 3D objects [24, 32], the data that is available is purely "virtual" and consists of either image-based data or 3D point clouds/polygon meshes. AR research often requires physical props (plus, ideally, a digital model of them), for example for blending of real and virtual content, user interaction, or re-lighting.

Researchers have been very creative in finding props for this purpose (cf. Figure 2 for a few examples), but we are not aware of any standard object or model that is available to and useful for many. Our assumption is that this is mostly due to the difficulty of distributing or replicating a physical object.

3 Designing an AR stage set

3.1 Applications

The major purpose of our model is to meet the needs and requirements of various AR applications. AR covers many different research fields such as computer vision, computer graphics, interface design and more. We designed the "City of Sights" with the following research areas in mind, and determined what kind of features, data and ground truth have to be provided to ensure the model's usefulness for the particular area:

Vision-based tracking, detection and recognition: Vision-based tracking and/or pose estimation is a basic requirement for many AR applications. There exist various approaches such as model-based tracking and detection, for example from natural features [20, 37], and simultaneous localization and mapping (SLAM) [15, 31]. A main requirement for the target scene is the existence of distinct natural feature points and structures (e.g. lines), and, for evaluation, varying degrees of this. Comparison with the camera pose ground truth is the most straightforward way to assess the correctness and quality of the algorithms.

Online/Real-time 3D modeling: For interactions between the real and virtual world that go beyond annotation, 3D models of the environment are needed (for example for occlusion management [19]). For AR, especially online (on-the-fly) 3D modeling is of interest [26]. Test objects have to satisfy the requirements for vision-based tracking and detection (see above) and consist of three dimensional objects of varying complexity (e.g. level of detail, number of planes, type of primitives). For evaluation, camera pose ground truth and accurate measurements of the target objects itself are needed.

Spatial AR/Virtual (Re-)lighting/Visual Coherence: This area mainly deals with the re-lighting of objects and blending of real and virtual content in a seamless manner, see e.g. [2, 16, 28]. Because spatial AR and related applications require a virtual representation of the real world object as well a real world object that is appropriate for projection, white objects are commonly used (cf. bottom row of Figure 2). Objects with interchangeable textures and materials provide varying challenges to different compensation techniques. The setup should be both feasible and challenging for the placement of projectors and interaction with hand-held objects such as palettes and brushes.

Tele-collaboration: In tele-collaboration, AR "allows us to leverage the advantages from both the worlds" [9]. Especially for interaction (tangible tele-collaboration), both a physical and digital model has to exist [7], and for some tasks, two physical instances of the model. To support robotic telepresence, it would be advantageous to have models within reach of a robot arm.

User interface design, Visualization in AR: X-ray vision [3] and label/annotation placement strategies [5] are two commonly researched areas in visualization for AR. Test objects should have a defined and plausible interior for X-ray visualization, and provide 3D-referenced meta-information for labels, ideally enough in terms of complexity/hierarchy and quantity that naive display of all labels will lead to clutter and advanced placement strategies and information filtering are needed.

Simulation of AR: To provide better control over environment parameters and investigate effects of immersion factors that are not yet available through current AR hardware, AR may be simulated using an appropriate virtual reality setup [8, 18]. With exactly comparable physical and virtual models, validation experiments can be run which test some participants using a real AR system, and some with a simulated AR system. Such experiments could test the transfer of research results from simulation to real deployments.

3.2 Requirements

With the applications listed above in mind, we identified the following set of desirable properties for the model to be useful for a wide range of AR research:

(1) it should exist physically and virtually,
(2) it should exhibit a range of properties/complexities in terms of texture and geometry,
(3) it should be customizable and extensible,
(4) it should be physically replicable by anyone,
(5) it should be accompanied by ground-truth observations and meta-data.

Of these, especially (1) and (4) seem to be difficult to fulfill at the same time and rule out all purely image or video based datasets.

[1] http://graphics.stanford.edu/data/3Dscanrep/

3.3 Creation approaches

With the requirement that the model has to exist both digitally and physically, two creation approaches are possible: (a) start with a physical model and construct the digital model from it, or (b) start with a digital model and construct the physical model from it. For the first approach, there are several highly accurate reconstruction methods such as controlled multi-camera vision based methods or laser scanning. Although there is no limitation in the complexity of the model itself, the digitalization of the model can imply a limitation in the shape and accuracy of the model. However, it is rather complicated to ensure requirement (4), specifically, that a specific physical model with the desirable properties is available world-wide in exact replication. Moreover, with this approach customization (3) is very difficult to achieve.

The two main advantages of the second approach, namely, starting with the digital model, is that it is easy to distribute and easy to customize, for example to change parameters such as scale, texture, surface properties or geometry. In this case a physical representation has to be created. Several options for manufacturing a physical model based on a digital model come to mind. Two options which are reasonably cheap (compared to industrial manufacturing) are the usage of a 3D printer or the creation of paper models. The major advantage of 3D printed models is the accurate manufacturing process and the possibility of creating models with higher complexity. In comparison to that, paper models are much cheaper to produce (by a factor of ten to twenty), are easier to distribute and provide more customization possibilities e.g. by changing textures or geometry.

4 Our Approach

Using a 3D printer allows for high complexity and accuracy of the model, but requires special equipment and thus may not satisfy requirement (4). We therefore decided to use paper models, which come at the cost of limited accuracy and complexity, but have several practical advantages which make them a good fit with the desired requirements. In particular:

- with an appropriate workflow (described in Section 4.2), they are easy to design.
- they are easy to distribute (all you need is the folding plan) and reproduce: print, cut, assemble. No additional purchases or special equipment are needed (although we do recommend a sharp blade and a proper cutting mat!).
- they are highly customizable: it is easy to change the scale, exchange or remove textures (cf. Figure 3), add markers if needed by the application (cf. the second image in Figure 2), change surface properties (matte vs. glossy paper), and even change the geometry.

We suggest to complement this model with a 3D printed object or widely available industrially manufactured and scanned object to increase the variety in terms of level of detail and surface properties if needed. An example is given in Figure 3 where we added plastic toys for which digital models can be obtained.

4.1 Model selection

The "City of Sights" consists of models of the following buildings:

- the Pyramid of Cheops (also called the Pyramid of Khufu),
- the Berliner Dom (Berlin Cathedral),
- the Arc de Triomphe de l'Etoile in Paris,
- the Musikverein in Vienna (Vienna concert hall),
- a medieval Irish Round Tower,
- St. Mark's Campanile in Venice.

Figure 3: Top row: Two examples of customization (change in scale, removal of texture). The latter is interesting especially for virtual relighting, cf. Figure 2 bottom. Bottom row: adding other objects increases variety in terms of surface structure (the toy trees) and material properties (the reflection on the car's windshield).

The reason for using models of real monuments is availability of models and additional meaningful data (e.g. names and labels, cf. Figure 11(b)). No other relationship with the real monuments is intended, and we do not have data on comparisons to (measurements or observations of) the real buildings.

We intentionally chose objects of different geometric complexities, including different geometric primitives such as boxes, approximate cylinders (the Round Tower), domes (on the Berliner Dom), concave surfaces (most prominently in the Arc de Triomphe), and small details (the rims of St. Mark's Campanile) as well as a variety of textures ranging from complex to repetitive to low textured.

We limited the overall size and the number of objects so that the model at its default scale fits on a ground plane of 800x550 mm (slightly smaller than A1), and that the total time needed for construction remains manageable. We arranged the models so that both un-obstructed views of all buildings as well as views with occlusions (interesting for X-ray vision and label placement, cf. Section 3.1) are possible with the limited reach of a robot arm (cf. Section 5.1). The proposed arrangement may be seen in Figure 1. The object sizes were chosen such that all paper folding plans fit on A3 or 11x17" sheets. The absolute scales of the models are not matched — we considered the ease of printing (see above) and assembly to be more important.

4.2 Paper model design & construction

Our workflow for creating paper models is illustrated in Figure 4. The first step is to find or create 3D models of buildings and objects which can be adapted into a paper folding plan. We found suitable models in the Google Warehouse[2], which is a repository for models created by users of Google SketchUp and Google Earth. The buildings are usually modeled and textured from street-level images, along with satellite photos.

We edited the models using Google SketchUp to remove details which would not be possible to reproduce with paper models, and to clean up any extraneous geometry. Then, using the Pepakura software package[3], we optimized the folding and cutting plan. For some models, we went through several iterations of editing, printing and cutting to find a good folding plan: while Pepakura does a great job of "unfolding" the model and creating the correct folding plan geometry, we found it necessary to manually optimize the position of cuts and the position, shape and geometry of the attached flaps

[2] http://sketchup.google.com/3dwarehouse/
[3] http://www.tamasoft.co.jp/pepakura-en/

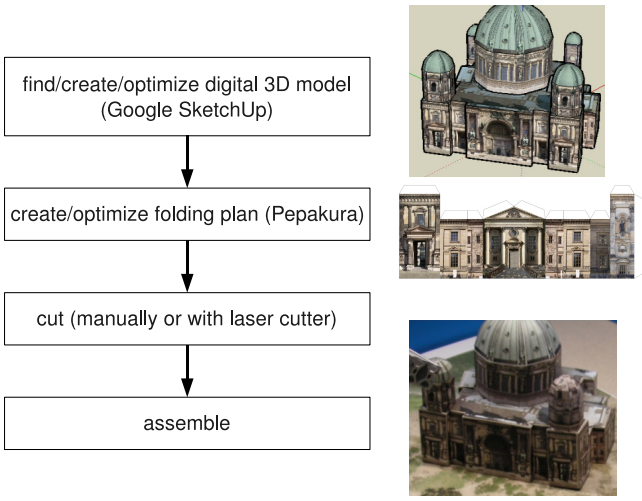


Figure 4: Workflow to create paper models. Unless the user wishes to change geometric features, the amount of work to create a copy of our "City of Sights" is reduced to the last two stages: print and cut the folding plan and assemble the model.

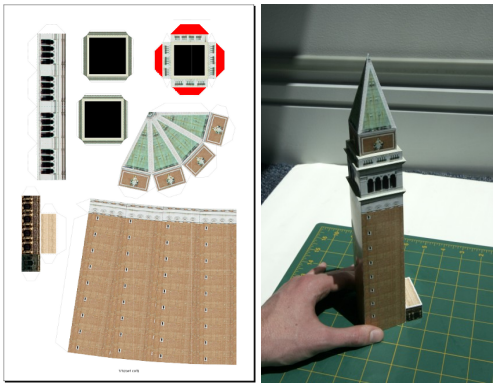


Figure 5: Folding plan and assembled model of the St. Mark's Campanile. On one piece, the folding flaps are highlighted in red.

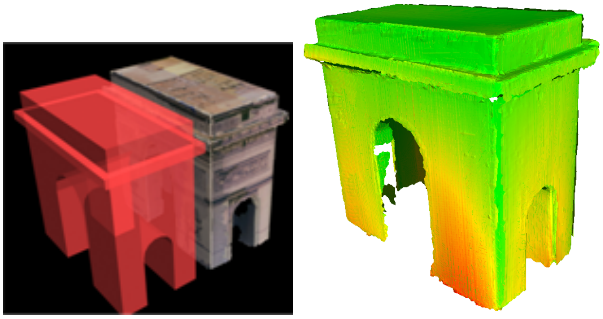


Figure 6: Left: digital source model (in transparent red) and result of the laser scan (textured) next to each other; right: difference between the two color-coded (right image, red for high error, green for low). Most of the model is accurate within 2-3 millimeters. The highest error occurs at the base which bends outwards due to the flexibility of the paper.

(highlighted red in Figure 5) in order to make the model easy to assemble and sufficiently rigid.

After preparing and constructing so many paper models, we developed some insights into the challenges involved and the right way to handle them. Some lessons learned are listed here: the side of the flap (if part A & B go together, does A or B have the flap?) can be very important; the ease of assembling all seams is much more important than the number of individual seams; small details are doable if the piece is well-designed; it is very advantageous if the part-in-assembly has flat sides (on which one may exert force while assembling) as long as possible. A well-planned folding layout and order of assembly is needed especially for the inwards facing surfaces (for example for the Arc de Triomphe) and small details such as the rims around St. Mark's Campanile (which are only 3 mm wide). We will provide specific assembly hints together with the models (cf. Section 6).

220 g/m^2 card stock paper was identified as a good trade-off between ease of cutting and assembling and sturdiness of the models. For larger flat faces as in the Musikhaus, T-shaped "paper beams" attached from inside help to stabilize the model. To further improve the rigidness of the models (especially if used in user interface experiments in which the user is expected to move them), we tested filling the models with hardening foam. With careful handling during the hardening process (e.g. supporting large faces), this improves the sturdiness considerably without change in visual appearance.

One sheet of the folding plan takes roughly one hour to cut, fold, and assemble, less for the box-shaped Musikhaus and more especially for the domes of the Berliner Dom. Each model consists of one (Campanile, Round Tower) to five (Berliner Dom) sheets. The whole model can be assembled from scratch in about 16 hours, less with some practice and more for especially meticulous assembling. Figure 5 shows an example of a paper folding plan and the assembled model.

4.3 Accuracy

To determine the overall accuracy that can be achieved with this workflow, we created a digital reconstruction of one of the paper models using a NextEngine 2020i laser scanner, which offers an accuracy of 0.12 mm at a resolution of 15.7 samples/mm. We chose to scan an early (and hence not too expertly assembled) instance of the Arc de Triomphe, as this model is a particularly difficult case in terms of accuracy: due to the arcs on each side, the base cannot be supported by a continuous piece and hence the four sides have a tendency to be pushed outwards by the arcs. Using MeshLab[4], we aligned the scanned mesh and the digital source model with iterative closest points and measured the Hausdorff distance (from scanned to source). Mean, root mean square, and maximum errors were 1.93 mm, 2.46 mm, and 9.58 mm respectively. The distribution of errors is visualized in Figure 6.

We conclude that our approach allows an accuracy of 2-3 millimeters for most parts. Distortions can be minimized by meticulous assembly and attaching rigid beams on the inside, but this is limited to a certain extent, and our workflow will not be a good fit for applications in which sub-millimeter accuracy is crucial. If higher accuracy is needed, 3D printed models or industrially manufactured objects should be used.

5 Acquisition of Video and Ground Truth Data

In the following we describe the video and ground truth data that we collected to accompany the "City of Sights". We used three different methods for camera control and ground truth acquisition: a Mitsubishi Rv14 robot arm, a manually guided and mechanically tracked Faro CMM arm, and an optical tracking system by ART.

All measurements used in the various calibration steps were recorded and are available together with the video and ground-truth data. This enables every researcher to both assess the quality of

[4] http://meshlab.sourceforge.net/

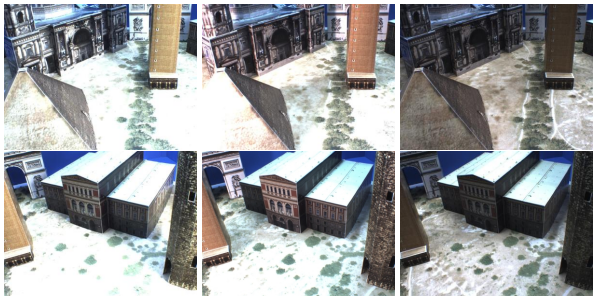


Figure 7: Example frames from video sequences recorded with the robot arm: two different viewpoints with three different lighting conditions each.

the sensor data and to employ different algorithms (e.g. for camera calibration) if desired.

5.1 Robot Arm Sequences

The computer-controlled robot arm was used to make "perfect" image sequences without the problems of handheld camera movements. It has the advantage that the exact same camera path can be repeated while changing properties of the environment such as models, lighting and camera parameters. These image sequences may be used in a variety of AR problems in which very accurate registration is needed, including reconstruction, visual coherence such as shadowing, occlusion handling or virtual relighting, but also to benchmark tracking algorithms under various lighting conditions.

The robot arm's position was moved in about 0.1 mm increments between frames, and we captured still images of 1600×1200 pixel resolution free of motion blur, frame drops or jitter. The resulting sequences can then be put together into a video sequence or treated as separate high-quality images.

We programmed three paths which move the camera by slowly panning, rotating, and moving closer or farther away from the buildings. The range and size of the robot arm, however, restricted movement around the model and also prevented any movements far into the interior of the space. All three paths were recorded with three different lighting conditions (illustrated in Figure 7). We measured the light intensity at three reference points for each light situation and registered the position and direction of the light sources in the scene.

5.2 Mechanically and Optically Tracked Sequences

We further used a mechanical and an optical tracking setup (Faro CMM and ART, respectively) to provide additional image data with ground truth. Here, the ground truth is less precise than for the robot arm, but the setup allows for direct user-controlled camera movements including real-world camera issues such as motion blur or jitter. The Ubitrack tracking framework [13, 27] was used for calibration of both setups as well for recording their sensor data.

Calibration of the Faro CMM. The Faro Fusion is a mechanical coordinate measurement machine (CMM) which measures the position and orientation of its tip in the base coordinate frame in real-time. This offers accurate and robust ground truth data, albeit with reduced maneuverability due to the bulkiness of the arm. For this setup, a Logitech Quickcam Pro 4000 was mounted rigidly to the Faro tip (see Figure 8).

A standard camera calibration of the Logitech camera was performed using the calibration algorithm by Zhang [38]. With the calibrated camera and a fixed chessboard-marker, the relationship between the base of the Faro CMM and the chessboard marker was determined using absolute orientation [12]. Then, the pose of the

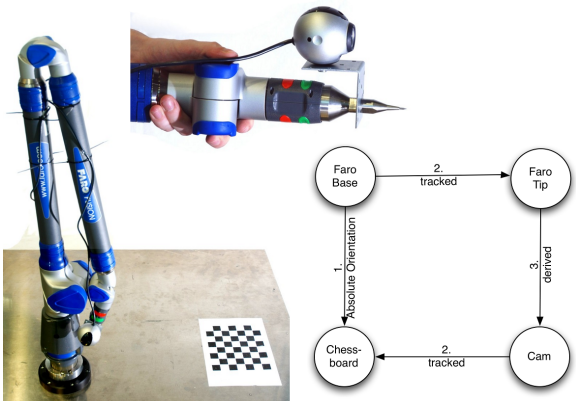


Figure 8: The Faro CCM arm, tip with attached camera, and the spatial relationship graph.

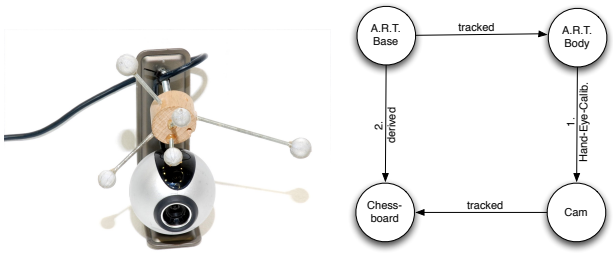


Figure 9: The camera with attached ART body and the spatial relationship for this setup.

chessboard marker was tracked in the undistorted camera image. These measurements can then be used to infer the spatial camera calibration, cf. Figure 8.

Using the tracking data from the chessboard detection, the temporal offset of images and sensor data was minimized by maximizing the normalized cross correlation between both signals [14, 17] while performing appropriate distinctive movements.

The video data was captured as a series of 640×480 still images to guarantee a 1:1 relationship between individual frames and timestamps. To avoid frame drops during the capture of the video images, the data was captured into a RAM drive and stored later.

Calibration of the ART Optical Tracker. The ART system is a multi-camera infrared optical rigid-body tracking system which uses constellations of retroreflective marker balls as bodies which can be attached to the object which is to be tracked. Compared to the Faro CMM, it offers less accurate tracking but does not impose any constraints on the user's camera movement. As for the Faro CMM, a Logitech Quickcam Pro 4000 was rigidly connected to an ART marker body (cf. Figure 9).

Most of the calibration steps were performed in the same way as for the Faro CMM setup. These steps include the camera calibration, temporal calibration and calibration of the stage to the ART coordinate frame, using the ART body and a calibrated ART measurement tip, respectively, instead of the Faro tip.

The major difference to the Faro setup calibration is the spatial calibration of the camera coordinate frame to the ART body, which in this case was computed directly using the hand-eye-calibration algorithm Tsai and Lenz [35], as illustrated in Figure 9. The spatial relation between the chessboard marker and the ART system (needed to verify the calibration) was calculated in a second step using tracked and calibrated data.

Figure 10: Two exemplary frames of the user-controlled sequences.

Recording of user-controlled video. With these more freely moving systems, we were interested in capturing different movement patterns as created by different types of users. We asked twelve participants (mainly non-computer vision experts) to generate a video sequence of the model using both the mechanical and the optical system. They were not given specific instructions about what to look for but were just told to explore the model. They were also not told details about the particular tracking system being used.

Each participant was given one minute to capture a video of the model. The participant could watch the live output of the camera on a nearby monitor, and could switch off the video feedback if they wished. This setup is similar to an AR application in which a video see-through display is used with indirect movement.

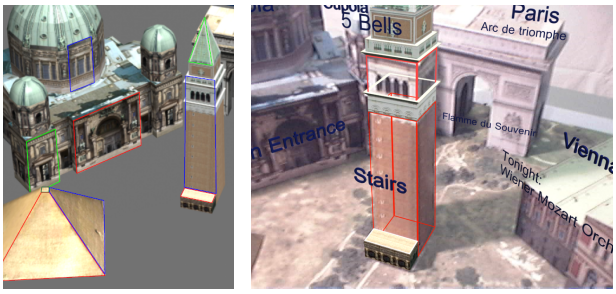
The video sequences produced from this experiment are quite different from the controlled robot arm sequences discussed above. Some example frames are shown in Figure 10. As to be expected, the sequences contain motion blur and jittery motion as a result of fast and unstable hand movements. Also, the sequences have discontinuities when the camera turns away from the model in an unexpected manner, such as when it is being repositioned or pointed at something else. These sequences taken "in the wild" by a variety of users, but with ground truth tracking, represent an interesting challenge for tracking systems and other AR technologies which need to be robust to the many issues inherent in handheld video sequences.

6 Data Set Summary

In summary, the initial dataset that is available on our website[5] includes the following items:

- A set of textured digital models (.skp, .3ds) and paper folding plans (Pepakura's file format .pdo and ready-to-print PDFs) for six buildings of varying geometric complexity,
- nine video sequences created with the robot arm consisting of three camera paths with three different scene lightings each (total of about 18 000 frames at 1600x1200 pixel),
- 12 video sequences created with the Faro CCM by non-experts (total of 11 039 frames at 640x480 pixel), including all measurements involved in creating the ground truth,
- 12 video sequences created with ART arm from non-experts (total of 11 598 frames at 640x480 pixel), including all measurements involved in creating the ground truth,
- world coordinates of 18 distinct planes of the real world model (see Figure 11(a)) measured with the Faro CCM arm, which may be used as ground truth reference observations for plane fitting/reconstruction algorithms,
- a set of labels registered to 3D points, surfaces and volumes in the scene which may be used to test visualization techniques. Figure 11(b) shows an AR view featuring a partial wireframe model and some of the labels.

[5] http://cityofsights.icg.tugraz.at/



(a) planes (b) labels

Figure 11: (a) Visualization of the reference planes, (b) AR view of the City of Sights, featuring one frame of the Faro-tracked videos augmented by a partially wireframe, partially textured model of the Campanile and some of the labels provided as meta-data.

It should be noted that the City of Sights is intentionally designed to be customizable and extensible. While we hope that above data will be helpful for many applications, we explicitly welcome extensions or variations and provide all intermediate data needed to derive those (cf. Section 5.2).

7 Conclusions

In this paper, we presented our approach for a standard model dataset which is useful as a "stage" for a variety of AR research. In contrast to other datasets, which consist only of image or video data, we make available the "blueprints" of the "City of Sights", enabling each researcher to create a physical copy of the model and/or extend the accompanying set of video sequences.

Our set consists mainly of paper models which we found to be a good fit for the desired properties derived from a set of potential applications. In particular, the model is easy to distribute, does not require any special equipment for assembling, and allows for very easy and comprehensive customization. Although paper models limit the accuracy and complexity of the set, they can be optionally supplemented with single differently-manufactured objects. Insights from the creation process include details on paper strength, what makes a folding plan a good folding plan, the smallest manufacturable level of detail, approximate assembly times, and the accuracy that may be achieved with this workflow.

We also provide several video sequences under varying conditions, and detailed the respective ground truth processes. We collected both computer-controlled and handheld video sequences which, in combination with the ground truth 3D model data, form a rich dataset for evaluation and testing of AR technologies and systems.

Acknowledgements

We want to thank Matthias Ruether and Martin Lenz (ICG TU Graz) for their assistance with the robot arm, and Cha Lee and Chris Coffin (UCSB) for testing the model stabilization with foam. This work was partially supported by the Christian Doppler Laboratory for Handheld Augmented Reality, the Austrian Science Fund FWF under contract W1209-N15, the Austrian Research Promotion Agency (FFG) under contract no. FIT-IT 820922, NSF CAREER grant IIS-0747520, ONR grant N00014-09-1-1113, and by the German BMBF project AVILUS (grant 01M09001V). We further want to thank the anonymous reviewers and especially the coordinator of our submission for comments and feedback.

REFERENCES

[1] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *Proc. Int'l Conf. on Computer Vision (ICCV)*, pp. 1–8, 2007.

[2] D. Bandyopadhyay, R. Raskar, and H. Fuchs. Dynamic shader lamps: painting on movable objects. In *Proc. IEEE/ACM Int'l Symposium on Augmented Reality*, pp. 207 –216, 2001.

[3] R. Bane and T. Höllerer. Interactive tools for virtual x-ray vision in mobile augmented reality. In *Proc. 3rd IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'04)*, pp. 231–239, Washington, DC, USA, 2004.

[4] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *Int'l Journal of Computer Vision*, 12(1):4377, 1994.

[5] B. Bell, S. Feiner, and T. Höllerer. View management for virtual and augmented reality. In *Proc. 14th ACM Symposium on User Interface Software and Technology (UIST'01)*, pp. 101–110, New York, NY, USA, 2001.

[6] G. Carneiro and D. Jepson. Flexible spatial models for grouping local image features. *Proc. 2004 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 747–754.

[7] H. Fuchs. Beyond the desktop metaphor: Toward more effective display, interaction, and telecollaboration in the office of the future via a multitude of sensors and displays. In *Proc. First Int'l Conference on Advanced Multimedia Content Processing (AMCP'98)*, pp. 30–43, London, UK, 1999.

[8] J. L. Gabbard, J. E. Swan, II, and D. Hix. The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality. *Presence: Teleoper. Virtual Environ.*, 15 (1):16–32, 2006.

[9] S. K. Ganapathy, A. Morde, and A. Agudelo. Tele-collaboration in parallel worlds. In *Proc. 2003 ACM SIGMM workshop on Experiential Telepresence (ETP'03)*, pp. 67–69, New York, NY, USA, 2003.

[10] S. Gauglitz, T. Höllerer, P. Krahwinkler, and J. Roßmann. A setup for evaluating detectors and descriptors for visual tracking. In *Proc. 8th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'09)*, pp. 185 –186, oct. 2009.

[11] L. Gruber, S. Zollmann, D. Wagner, T. Höllerer, and D. Schmalstieg. Optimization of target objects for natural feature tracking. In *Proc. IEEE Int'l Conference on Pattern Recognition 2010*, 2010.

[12] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 4(4), Apr. 1987.

[13] M. Huber, D. Pustka, P. Keitler, E. Florian, and G. Klinker. A system architecture for ubiquitous tracking environments. In *Proc. 6th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nov. 2007.

[14] M. Huber, S. Michael, and G. Klinker. Temporal calibration in multisensor tracking setups. In *Proc. 8th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'09)*, Orlando, USA, October 2009.

[15] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. 6th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[16] G. Klein and D. Murray. Compositing for small cameras. In *Proc. 7th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'08)*, pp. 57–60, Washington, DC, USA, 2008.

[17] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics Speech and Signal Processing*, 24(4):320–327, 1976.

[18] C. Lee, S. Bonebrake, T. Höllerer, and D. A. Bowman. The role of latency in the validity of AR simulation. In *Proc. 2010 IEEE Virtual Reality Conference*, pp. 11 –18, 20-24 2010.

[19] V. Lepetit and M.-O. Berger. A semi-automatic method for resolving occlusion in augmented reality. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*, vol. 2, pp. 225 –230 vol.2, 2000.

[20] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(9): 1465–1479, 2006.

[21] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In *Proc. 8th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'09)*, 2009.

[22] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int'l Journal of Computer Vision*, 60(1):63–86, 2004.

[23] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, Oct. 2005.

[24] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *Int'l Journal of Computer Vision*, 73 (3):263–284, 2007.

[25] M. E. Newell. *The utilization of procedure models in digital image synthesis.* PhD thesis, 1975.

[26] Q. Pan, G. Reitmayr, and T. Drummond. ProFORMA: Probabilistic Feature-based On-line Rapid Model Acquisition. In *Proc. 20th British Machine Vision Conference (BMVC)*, London, September 2009.

[27] D. Pustka, M. Huber, M. Bauer, and G. Klinker. Spatial relationship patterns: Elements of reusable tracking and calibration systems. In *Proc. 5th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'06)*, October 2006.

[28] R. Raskar, G. Welch, K.-L. Low, and D. Bandyopadhyay. Shader lamps: Animating real objects with image-based illumination. In *Proc. 12th Eurographics Workshop on Rendering Techniques*, pp. 89–102, London, UK, 2001.

[29] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Int'l Journal of Computer Vision*, 37(2):151172, 2000.

[30] J. Schöning, M. Rohs, and A. Krüger. Mobile interaction with the "real world". In *Proc. Mobile HCI 2008: Workshop on Mobile Interaction with the Real World (MIRW)*, 2008.

[31] G. Schweighofer, S. Segvic, and A. Pinz. Online/realtime structure and motion for general camera models. In *Proc. 2008 IEEE Workshop on Applications of Computer Vision (WACV'08)*, pp. 1–6, Washington, DC, USA, 2008.

[32] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:519–528, 2006.

[33] H. Tamura and H. Kato. Proposal of int'l voluntary activities on establishing benchmark test schemes for ar/mr geometric registration and tracking methods. In *Proc. 8th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'09)*, pp. 233 –236, oct. 2009.

[34] A. Torrence. Martin newell's original teapot. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Teapot*, p. 29, New York, NY, USA, 2006.

[35] R. Y. Tsai and R. K. Lenz. A new technique for fully autonomous and efficient 3D robotics hand-eye calibration. *IEEE Journal of Robotics and Automation*, 5(3):345–358, June 1989.

[36] G. Turk and M. Levoy. Zippered polygon meshes from range images. In *Proc. 21st Annual Conference on Computer graphics and interactive techniques (SIGGRAPH'94)*, pp. 311–318, New York, NY, USA, 1994.

[37] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg. Pose tracking from natural features on mobile phones. In *Proc. 7th IEEE/ACM Int'l Symposium on Mixed and Augmented Reality (ISMAR'08)*, Cambridge, UK, Sept. 15–18 2008.

[38] Z. Zhang. A flexible new technique for camera calibration. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[39] K. Zimmermann, J. Matas, and T. Svoboda. Tracking by an optimal sequence of linear predictors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31:677–692, 2008.