Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps

Sara Martull
University of Tsukuba, Japan
E-mail: info@martull.com

Martin Peris
Cyberdyne Inc., Japan
E-mail: martin_peris@cyberdyne.jp

Kazuhiro Fukui
University of Tsukuba, Japan
E-mail: kfukui@cs.tsukuba.ac.jp

Abstract

Stereo matching is one of the most active research areas in computer vision. While a large number of algorithms for stereo correspondence have been developed, research in some branches of the field has been constrained due to the few number of stereo datasets with ground truth disparity maps available. Having available a large dataset of stereo images with ground truth disparity maps would boost the research on new stereo matching methods, for example, methods based on machine learning. In this work we develop a large stereo dataset with ground truth disparity maps using highly realistic computer graphic techniques. We also apply some of the most common stereo matching techniques to our dataset to confirm that our highly realistic CG stereo images remain as challenging as real-world stereo images. This dataset will also be of great use for camera tracking algorithms, because we provide the exact camera position and rotation in every frame.

1. Introduction

Stereo vision is a very active research topic, every year several new stereo matching methods are introduced [5]. The goal of these stereo matching algorithms is to accurately generate a dense disparity map, which describes the difference in location of corresponding features seen by the left and right cameras. To measure and compare the performance of such algorithms it is essential that the ground truth disparity map is known.

Several stereo datasets with known ground truth disparity maps are available [5, 6], but the number of stereo pairs is very limited. This limitation has been constraining the progress of research in some branches of the field. For example, trying to apply machine learning (ML) techniques to solve the stereo matching problem has been very difficult due to the fact that ML usually requires large amounts of data with ground truth for learning and very few are available. Some efforts on using Computer Graphics (CG) synthetic data have been made [1], but the simplicity of the generated scenes makes the stereo matching problem unrealistically easy to solve.

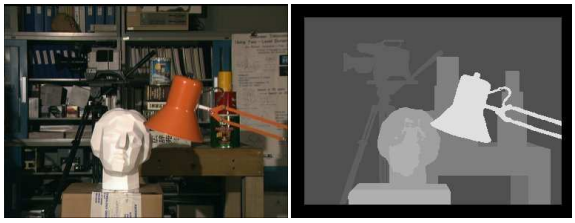

Figure 1: *Head and Lamp* scene. Left image and ground truth disparity map.

Among the available datasets, one of the most known and extended scenes is the *Head and Lamp* (Figure 1) stereo dataset developed at University of Tsukuba [3].

In this work we created a highly realistic CG dataset that properly models real-world imperfections, while providing accurate ground truth. It is based in the original *Head and Lamp* set of images, as a tribute to the early efforts of the University of Tsukuba in stereo vision, giving the chance to appreciate the scene from points of view not seen until now.

In addition, this dataset will be very useful for camera tracking algorithms, since we can provide the 3D position and rotation of the camera in every frame of the sequence.

Since we are working in a 3D environment, we can create any possible camera setting, and modify any camera parameter just as we would do with a real-world camera. In future datasets we will include video se-

Figure 2: We changed the illumination conditions on the CG scene to achieve new and more challenging datasets. Upper-left: Lamps. Upper-right: Fluorescent. Lower-left: Daylight. Lower-right: Flashlight.

quences with motion blur, lens aberrations, noise, and different shake effects, etc...

We contribute a dataset with the following properties:

- Different illumination conditions (Figure 2).
- 1800 full-color stereo pairs per illumination condition with ground truth disparity maps (1 minute video at 30FPS using an animated stereo camera).
- 256 levels of disparity.
- Non-occluded area mask, near depth discontinuity masks and 3D camera position and orientation on each frame.

The usefulness of this new dataset is confirmed on the work of Peris *et al.*([4]) where the CG data is used to train a ML-based stereo matching method and successfully applied to approximate the disparity map of a real-wolrd stereo pair.

2. Modeling and rendering process

The scene was generated using the software Autodesk Maya 2012. The original stereo pair from University of Tsukuba was used as image planes and all the geometry of the scene was then modeled using them as reference. Figure 3 shows an example of an early stage of the modeling process.

All the objects were modeled inside Autodesk Maya by poly modeling. We also created textures using reference photographs of real objects, and applied them to

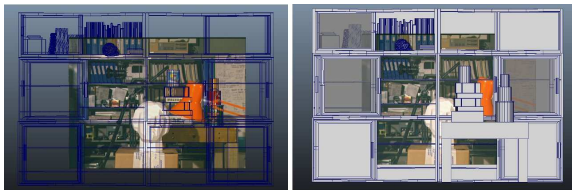

Figure 3: Geometry of the scene over the original image used as reference.

the 3D models in order to increase the amount of detail and realism.

Once all the geometry was modeled and adjusted to resemble the original scene, we added shaders and materials to every object in the scene, trying to replicate as accurately as possible the real world behaviors of the surfaces. Then, we added the virtual cameras and lights and rendered the scene, achieving the final video.

We generated 4 different versions of the video dataset by simply modifying the virtual lights in the scene into different illumination conditions: Daylight, Fluorescent lighting, Flashlight and Desk lamps (Figure 2).

Each illumination condition offers specific challenges that will be of great use for the improvement of new stereo and tracking algorithms:

- *Fluorescent*: This is considered the default illumination. It's lighting condition is perfectly even in the surfaces, and all the objects appear properly lit without exaggerated contrast between light and shadow.
- *Daylight*: This lighting condition gives a smooth illumination to the objects in the scene with exception of the areas near the window, that appear over exposed due to the intensity of the sun light.
- *Flashlight*: This scene has been rendered with anenvironmentin penumbra, only lit by the light of a flashlight attached to the moving Stereo Camera.The illumination is very low but even, leaving most of the scene in darkness with exception of the areas where the camera is pointing at.
- *Lamps*: This set the most challenging of all four illumination conditions, because it presents highly under exposed images, with low and uneven lighting.

3. Generation of ground truth data

The aim of this dataset is to provide ground truth data that can be useful to evaluate the performance of computer vision algorithms, especially for stereo vision and

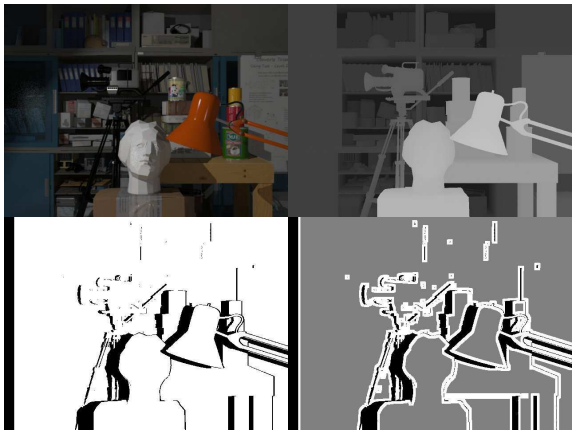

Figure 4: First frame of the new dataset (left camera view). Upper-left: RGB image. Upper-right: disparity map. Lower-left: non-occlusion mask. Lower-right: near depth discontinuity masl.

camera tracking methods. This section introduces the different kinds of ground truth data that we generated for this dataset. Figure 4 depicts some of the ground truth data available with this dataset.

3.1. Disparity Map

The ground truth disparity maps are useful to evaluate the overall quality of stereo matching algorithms.

After modeling the scene, the output of the render engine provides for each frame two RGB images and two grayscale images. The grayscale images were created using a custom ramp shader and represent the depth map of the scene (Figure 5). This depth map can be easily transformed into a disparity map knowing the camera intrinsic parameters (established in the camera properties of the CG software) and the following formula for each pixel on the depth image:

$$d = f\frac{T}{Z} \tag{1}$$

Where d is the disparity of the pixel, f is the focal length of the camera, T is the baseline distance of the stereo rig and Z is the depth value of the pixel.

However, there is a small limitation to this approach: the resolution of the rendered depth map is only 8 bits. This means that there would be only 256 possible values to represent the depth, which becomes quickly insufficient to represent all the depth values on a large scene. This in turn would cause the disparity map to be inaccurate.

We solved this issue by iteratively rendering the depth map changing the values of the near and far clipping planes so the distance between both planes is small

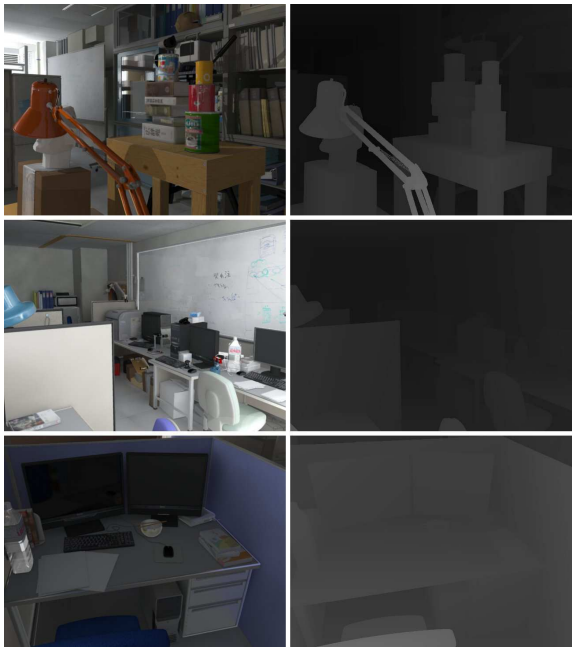

Figure 5: Some example frames from the dataset with their respective depth maps obtained by applying a custom *ramp shader* to the geometry of the scene.

enough to have a good depth resolution with only 256 values. Then later on, we obtain a single disparity map using the formula above and all the depth maps with different configurations of clipping planes.

The method of iteratively render portions of depth map allows us to generate ground truth disparity maps of arbitrary resolution. In this work we generated disparity maps with pixel accuracy (subpixel accuracy is left for further works).

We provide disparity maps for left and right cameras. See Figure 4.

3.2. Non-occluded area mask

In addition to disparity maps, for stereo matching method evaluation it is interesting to have a non-occluded area mask. This mask represents in white color the pixels on the scene that are visible from both cameras and in black color the pixels that are visible from only one camera.

To obtain the non-occluded area mask, we simply cross-checked the left and right disparity maps. Pixels that are visible in both cameras will have the same value in both disparity maps, but for occluded pixels the left and right disparity value will be different.

The performance of the stereo matching algorithm on areas where pixels are occluded is one of the most important quality indicators of the algorithm, as it is

very difficult to find the matching point of a pixel in one of the images if it is not visible on the other image.

We provide non-occluded area masks for left and right cameras. See Figure 4.

3.3. Near depth discontinuity area mask

Together with disparity maps and non-occluded area masks, near depth discontinuity area masks are the third kind of ground truth data used for standardized stereo matching algorithm evaluation. This kind of mask allows to evaluate the performance of the algorithms in areas near to a depth discontinuity (i.e. the boundary of an object). This is specially interesting as most methods tend to perform poorly in these areas.

To obtain this kind of mask we detected the boundaries of the non-occluded mask. In this way, a gray value represents a pixel visible from both cameras, a black pixels represents an occluded pixel and white pixels represent pixels that are near the transition between occluded/non-occluded pixels.

We provide near depth discontinuity are masks for left and right cameras. See Figure 4.

3.4. Camera position and orientation

In the image sequence that we generated, the camera is animated along a path. Using camera tracking methods this path should be able to be reconstructed. To evaluate the performance of such methods, we provide the 3D position and orientation of the camera on each frame.

To obtain this kind of ground truth data we executed a MEL script at render time that stored the 3D position and orientation of the camera on each frame.

We provide the position and orientation of the camera relative to the middle point of the stereo rig.

4. Experiments

We apply a well known stereo matching algorithm to our dataset and to the original dataset. The algorithm is based on Block Matching [2] and implemented in OpenCV library [7]. We will use the default parameters of OpenCV in our experiments. Figure 6 shows the output of the Block Matching algorithm for both, the original and the new, stereo pairs.

If we compare the results in Figure 6, we can observe that our CG dataset is even more complex than the original, as the original dataset has only 16 different disparity values (which is relatively easy to solve for stereo matching algorithms) and each stereo pair on our CG dataset has up to 256 possible disparity values making it more challenging to find the stereo correspondences.

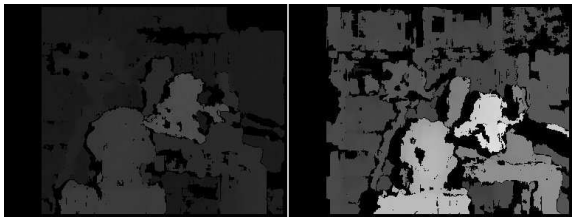

Figure 6: Output of Block Matching algorithm. Left: Original stereo pair. Right: CG stereo pair.

5. Conclusion

In this work we developed a highly realistic CG stereo dataset with several kinds of ground truth information that can be applied to stereo matching and camera tracking problems. We release it in the hope that it will be useful to further the research in the field of Computer Vision. The dataset can be found at: http://cvlab.cs.tsukuba.ac.jp

6. Acknowledge

This work was supported by KAKENHI (23650081).

References

[1] T. Frhlinghaus, J. M. Buhmann, and R. F. wilhelms universitat. Regularizing phase-based stereo. In *Proc. of ICPR*, pages 451–455, 1996.

[2] K. Konolige. Small Vision Systems: Hardware and Implementation. In *8th International Symposium on Robotics Research (ISRR)*, Oct. 1997.

[3] Y. Nakamura, T. Matsuura, K. Satoh, and Y. Ohta. Occlusion detectable stereo – occlusion patterns in camera matrix. In *CVPR*, pages 371–378, 1996.

[4] M. Peris, A. Maki, S. Martull, Y. Ohkawa, and K. Fukui. Towards a simulation driven stereo vision system. In *21st International Conference on Pattern Recognition*, Nov. 2012.

[5] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 47:7–42, 2001.

[6] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition*, CVPR'03, pages 195–202, 2003.

[7] WillowGarage. Opencv library, Jan. 2012.