Using Two Lenses for Depth Estimation and Simulation of Low Depth-of-Field Lenses

Andy L. Lin

Introduction

Recently, there has been a 3-D camera craze in the photography and video industry. For example, the recently launched Fujifilm W1 uses two lenses to capture two shifted images. These images are then used in 3-D displays to simulate depth, often with the help of polarized glasses. There are more capabilities that multi-lens cameras can offer, other than just capturing 3-D scenes.

One of such possibilities that has yet to be explored in the literature, is using inexpensive 2-lens cameras to simulate the artistic effect of low depth-of-field (DOF) lenses. These low DOF images which maintains focus on a small area of the scene, and blurs out the background are often the hallmark of professional and artistic photography. The way these lenses achieve low DOF is by using large apertures. Unfortunately, these low DOF lenses are very expensive themselves and can only be used in expensive digital single lens reflex (DSLR) cameras. One viable alternative is to use inexpensive 2-lens cameras to simulate the out of focus blurring effect of high-end lenses and cameras.

By using the parallax information in the two images captured by the single camera, a machine learning algorithm can estimate a depth map for the scene by estimating the disparity between the two images. With the proper depth information, an algorithm can then apply a lens blur function that varies with the depth of the scene, simulating the effect of an expensive low DOF lenses with large apertures.

Prior Work

Extensive amounts of previous studies have been dedicated to recovering depth maps from single images, as well as multiple images. For example, Saxena et. al [1] uses a combination of monocular and binocular cues to recover a depth estimation using the help of machine learning algorithms. A collection of state-of-the-art techniques and algorithms is described and documented in [2].

Since we are using a system with 2 lenses, it is sensible to focus on binocular stereo reconstruction algorithms. The simplest of such algorithms is described in [3]. For every point in one of the binocular pair images, the simple algorithm uses a sliding window approach to find the corresponding shifted window in the other image which yields a minimum sum-of-absolute-differences error. This distance is then a good estimate of the shifted location that yields the maximum correlation and thus a good estimate of the disparity between the two images at a specific pixel.

Simulating lens-blur functions depending on depth is also a well-studied problem. These blur functions have been experimentally determined and there are also simple equations that can be used to completely determine the type and amount of blur of a lens based on the focal length and the distance of interest in a scene.

Some simple image-editing software suites have implemented simple filters that attempt to reproduce the artistic effect of low depth-of-field. However, these filters are usually not adaptive to the scene of interest. For example, one simple filter is a progressively blurring filter that leaves the center of the scene in focus and applies a stronger blurring function as we approach the edges of the images. While this type of filter can produce images have only portions of the scene “in focus”, this type of filter cannot produce the sharp differences in focus and defocus necessary for artistic photography. While image segmentation algorithms may be able to solve this problem, they are often unreliable and often make mistakes.

The Stereo Reconstruction Algorithm

Classical Techniques

The sum-of-absolute-differences method is a simple way to obtain a depth estimate from a binocular pair [3]. Although computationally inexpensive and relatively easy to implement, this algorithm has a few weaknesses. For example, in uniform-texture regions of the image, it is difficult for the algorithm to find differences in the sum-of-absolute-differences at different apparent disparities. Thus, this baseline image makes many errors in uniform-texture regions [1].

Another weakness of the algorithm is trade-offs in window sizes. For small window sizes, the algorithm is susceptible to the noise in the image, so the resulting depth map will also be noisy. For large window sizes, the algorithm is less susceptible to noise; however, it is unable to properly calculate depth at depth map boundaries since each window will contain multiple disparities and attempt to calculate one value for multiple depths in the scene.

A simple way to improve upon this simple sum-of-absolute-differences algorithm is to apply a simple median filter on the depth map. Median filters are able to remove noise without blurring decision edges. Thus, applying a median filter to a depth map obtained by a small window-size can reduce noise without degrading the performance of the algorithm near decision edges.

Another way to improve upon the sum-of-differences algorithm is by employing a bilateral filter on the window. Bilateral filters are known to be able to blur uniform regions of images, but leave the edges of the image intact [4]. By applying a bilateral weighting function to the window depending on the pixel values of that window, we can then obtain depth maps which perform better near edges, but still retains performance elsewhere.

Machine Learning Techniques

To address these two main weaknesses, we can employ machine learning algorithms. Saxena et al. [1], uses a Markov Random Field model to combine depth maps obtained by different methods, as well as to enforce a continuity constraint. This method used is very effective in obtaining accurate depth maps from existing depth maps of different sources. Moreover, the algorithm is able to generate these depth maps that are robust to noise, particularly in uniform regions, due to the clever continuity constraint. However, one drawback of this approach is that it requires linear programming, which is somewhat computationally intensive.

Often times, a photographer wishes to receive real-time feedback on the photograph he/she has just taken so he/she can choose to take another photograph if the current one is not satisfying. The Markov Random Field method requires linear programming, so may take some time to process. Thus, a simpler approach was chosen for the project in order to reduce computational complexity to provide immediate feedback for the photographer. There are several different classical approaches to depth estimation. A simple way to improve upon these existing approaches is to use different approaches for different parts of the scene. Machine learning is used to determine which model to use, at specific regions of the scene. For example, in areas that are close to depth-map edges, it is desirable to use a small window-size scheme in order to take advantage of its behavior near object edges. On the other hand, it is desirable to use a large-window-size scheme in areas that are far from object edges in order to reduce noise.

Any feature that could be extracted that aids in determining which optimal model to use is useful. Other features used in the study included: image uniformity, sum-of-absolute differences error in optimal shift position, vertical position in image, estimated depth of scene.

As mentioned before, small window-size sum-of-difference algorithms are unable to properly estimate depth in areas of the image which are uniform. By choosing uniformity of the image as a feature allows the algorithm to avoid using small window-sizes for these regions. The minimum sum-of-absolute-differences error at each point is a feature similar to distance from object edge which can help determine if a specific classical method is working well at a particular point. Vertical distance and estimated depth of scene are features (taken from one of the classical methods), which both determine rough distance of the scene is meant to take advantage of the fact that object further from the camera tend to include denser details per pixel. Perhaps a smaller window size is needed for areas of the scene which are far away.

Experimental Setup

Data

The Middlebury Stereo project database was used for training and test sets. The database consisted of left and right images as well as ground truth images for various scenes. Ground truth images were obtained using structured light as described in [5]. For the study, 4 random binocular pairs were chosen for the training set, and 19 binocular pairs were used for the test set.

Comparisons

The sum-of-absolute-differences sliding window approach was the baseline for comparison [3]. As mentioned before, this algorithm is susceptible to trade-offs in window size. This median filter approach was the second algorithm used for the comparison. Another method to deal with window-size trade-offs is applying a bilateral weighting function on the window. This bilateral filter approach was the third algorithm used for comparison.

The possible conventional methods used for machine learning selection consisted of: 5x5 sum-of-absolute-differences, 9x9 sum-of-absolute-differences, 13x13 sum-of-absolute differences, 17x17 sum-of-absolute differences, median filter on 5x5 sum-of-absolute-differences, blurred median filter on 5x5 sum-of-absolute-differences, and 21x21 bilateral filter. The reason why the blurred median filter on 5x5 sum-of-absolute differences method was included was to provide a means for the algorithm to generate smoother data. This smoothed data is necessary because there is no formal smoothness constraint. The hope is for the algorithm to choose this model in areas of large errors, in efforts to keep the final depth map as smooth as possible.

Two types of machine learning classifiers were used for determining which model to use: the naïve Bayes classifier and logistic regression. The classical method that resulted in the best estimation of depth at a specific point was treated as the ground truth for that point.

Mean-square-error (MSE) for the depth maps was chosen as the error criteria. It is most effective to evaluate the error of these images by evaluating the error from the obtained depth maps. The reason for this is that the final blurred image will be completely dependent on these depth maps. Evaluating error on the final blurred image makes it more difficult to determine difference in quality.

Results

The new machine learning technique was able to outperform all conventional algorithms on the dataset used for the study. A preliminary trial of logistic regression versus the naïve Bayes technique was performed.

The naïve Bayes technique was able to outperform the logistic regression method, which also took much longer to evaluate. Thus, the naïve Bayes approach was used for the remainder of the study. As illustrated in Figure 2, the new machine learning technique consistently outperforms the conventional methods. On average, there is a 15% improvement in performance over the Median Filter Technique. See Figure 1 for the blurred images produced by these depth maps.

Though the obtained results show a great improvement in MSE compared to conventional methods, there are still improvements to be made. Most importantly, the estimated depth map contains sharp changes in depth that do not exist, even though some efforts were made to reduce these effects. These sharp changes in depth cause artificial artifacts in the final blurred image. Perhaps more stringent conditions on continuity should be used to prevent these artifacts from appearing in the future.

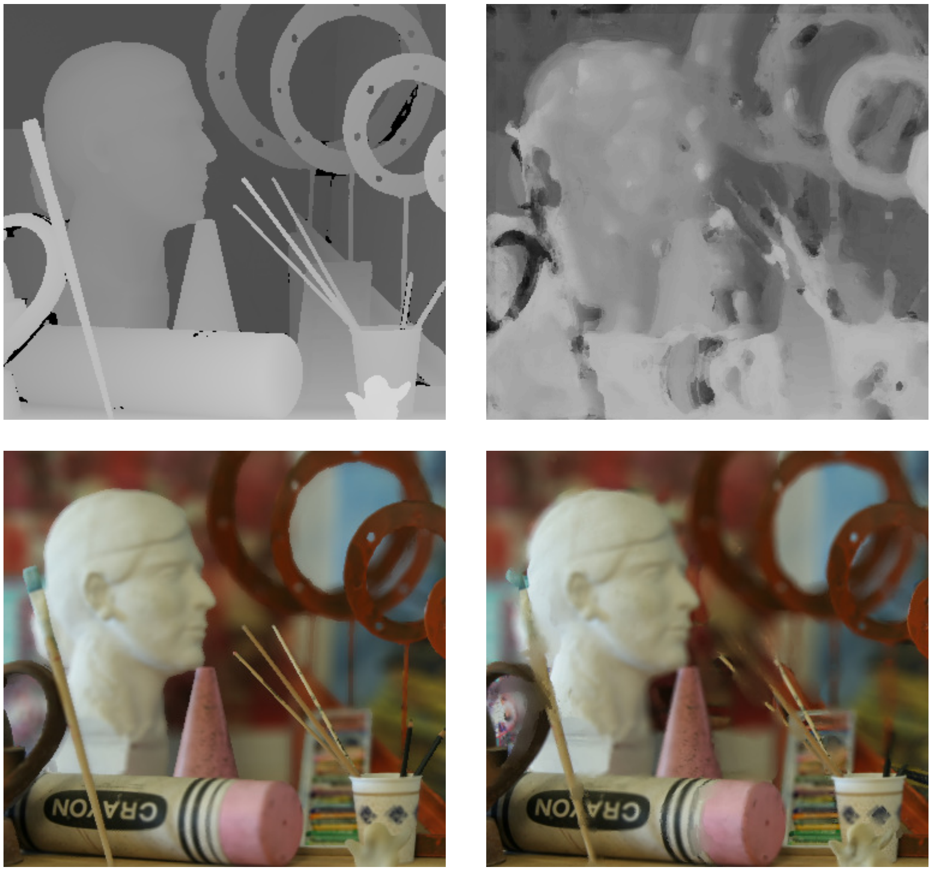

Figure 1: Ground truth depth map (upper-left). Estimated depth map (upper-right). Blurred image generated with estimated depth map (lower-right). Blurred image generated with ground truth depth map (lower-left). Blurred image generated with estimated depth-map (lower-right)

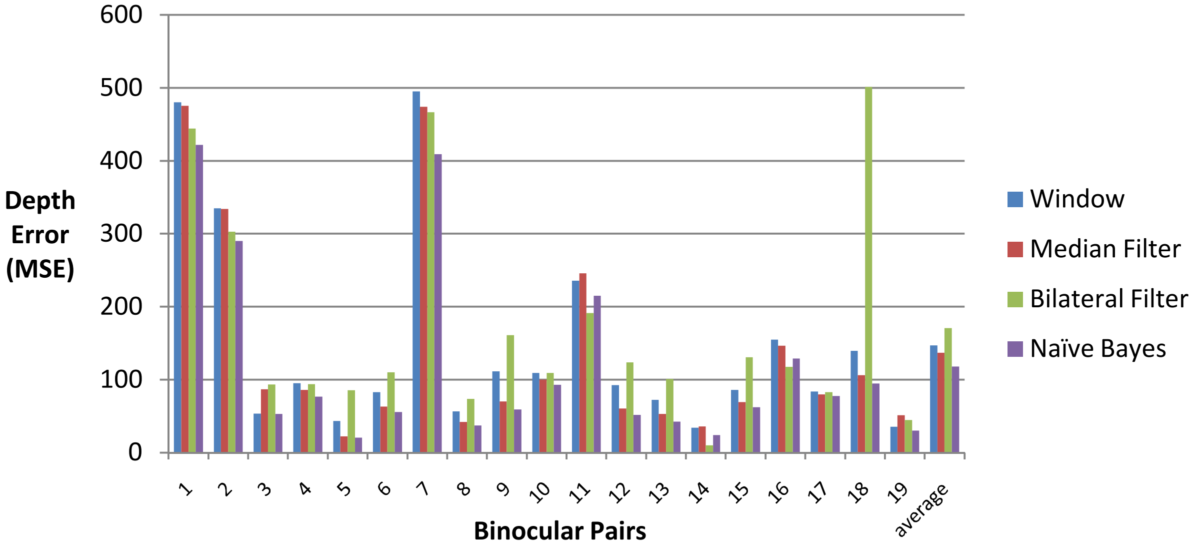

Figure 2: MSE comparisons of depth-map produced by different methods. The horizontal axis represents different test data (the average values are shown on the right).

Conclusion

With the help of machine learning, an algorithm which improves upon conventional techniques for depth estimation of binocular scenes was obtained. The machine learning technique consists of using features such as distance from decision boundaries and image uniformity to help determine which conventional depth-estimation algorithm to use in different areas of the image. The final result is depth maps which have less error than the conventional depth-estimation results. When this depth map is used as the basis for a lens-blurring function, plausible images that appear to be produced by large aperture lenses can be produced.

Acknowledgements

Thanks to Professor Andrew Ng and the CS229 staff for such an inspirational quarter. Also, thank you to Professor Brian Wandell for providing inspiration and advising me this quarter.

References

[1] A. Saxena, J. Schulte and A. Ng. Depth Estimation using Monocular and Stereo cues. *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 2007.

[2] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3):7-42, April-June 2002.

[3] David A. Forsyth and Jean Ponce, "Computer Vision: A Modern Approach", Prentice Hall, 2003.

[4] Volker Aurich and J¨org Weule. Non-linear gaussian filters performing edge preserving diffusion. *Proceedings of the DAGM Symposium*, 1995.

[5] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003),* volume 1, pages 195-202, June 2003.