IMPROVED IMAGE SELECTION FOR FOCUS STACKING IN DIGITAL PHOTOGRAPHY

David Choi, Aliya Pazylbekova, Wuhan Zhou, and Peter van Beek

Cheriton School of Computer Science, University of Waterloo, Canada

ABSTRACT

Focus stacking, or all-in-focus imaging, is a technique for achieving larger depth of field in an image by fusing images acquired at different focusing distances. Minimizing the set of images to fuse, while ensuring that the resulting fused image is all-in-focus, is important in order to avoid long image acquisition and post-processing times. Recently, an end-to-end system for focus stacking has been proposed that automatically selects images to acquire. The system is adaptive to the scene being imaged and shows excellent performance on a mobile device, where the lens has a short focal length and fixed aperture, and few images need to be selected. However, with longer focal lengths, variable apertures, and more selected images (as exists with other cameras, notably DSLRs), classification and algorithmic inaccuracies become apparent. In this paper, we propose improvements to previous work that remove these limitations, and show on eight real scenes that overall our techniques lead to improved accuracy while reducing the number of required images.

***Index Terms*—** Focus stacking, increased depth of field, computational photography

1. INTRODUCTION

Focus stacking combines several images captured at different focusing distances into a single image to produce a larger depth of field. It is useful when the camera is unable to acquire an all-in-focus image or when the quality of an all-in-focus image would be degraded because the narrow aperture results in a shutter speed too slow to freeze motion. The motivations for obtaining an all-in-focus image range from the aesthetic to the practical: architectural, interior, and macro photography, as well as pattern recognition and object detection [2, 3].

An essential part of focus stacking is selecting the set of images to be fused. The set must be small, in order to decrease capture and fusion times, but must result in an all-in-focus image. Time between image captures can be on the order of seconds, so even gradual motion can impact quality. However, in contrast to a wide literature on combining a set of images into a single image (see, e.g., [4–6]), image selection has not received much attention. The simple approach of moving the lens a uniform step-size and acquiring an image at each step leads to sets which contain images with nothing in focus or redundant images. Hasinoff et al. [7–9] considered the problem of quickly selecting the set of images to cover a given depth of field. However, their analysis neglects camera overhead, image post-processing, and ranges without objects.

Vaquero et al. [1] presented an end-to-end system that adaptively selects a minimal set of high-resolution images to acquire by processing a stream of low-resolution ones (which can be acquired quickly). It had the camera display a final all-in-focus image, allowing the photographer to verify the final result in the field (see Fig. 1). Their system showed excellent performance on a mobile device, where the lens has a short focal length and a fixed aperture.

Unfortunately, as we show, their techniques which work well for mobile devices do not necessarily generalize well to DSLRs, which feature lenses with longer focal lengths and variable apertures, and so require many more images for focus stacking. We propose improvements to the work of Vaquero et al. [1] which afford increased performance on non-mobile devices. Our improvements make use of shape from focus techniques (see, e.g., [2, 10, 11]), supervised machine learning techniques (see, e.g., [12, 13]), and standard depth of field equations to improve on previous inefficiencies. Empirically, on eight real scenes and various aperture settings, our techniques lead to an overall improved accuracy while significantly reducing the cardinality of the selected set of images.

2. OUR PROPOSALS

We first summarize Vaquero et al.'s [1] system (see Fig. 1), then describe our proposed improvements.

Vaquero et al.'s [1] approach proceeds as follows:

Step 1. Capture a stack of low-resolution images $p = 0, \ldots, n-1$ by sweeping the lens slowly enough that the depths of field for adjacent images overlap.

Step 2. Overlay a grid on each image and calculate a focus measure $\phi_{i,j}(p)$ for each cell (i, j) in the grid for each image (see Fig. 2). A focus measure maps an image to a value that represents its degree of focus (see, e.g., [14–16]). Let $f(x, y)$ be the luminance at pixel (x, y) in an image. Here, the focus measure for a cell of size $w \times h$ pixels is given by:

$$\phi_{i,j}(p) = \sum_{x=0}^{h-1} \sum_{y=1}^{w-2} \mid -f(x, y-1) + 2f(x, y) - f(x, y+1) \mid .$$

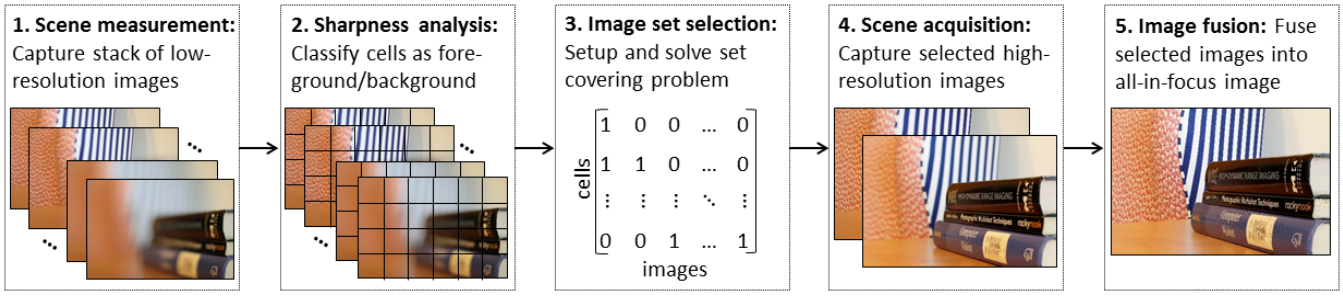

Fig. 1. Pipeline for Vaquero et al.'s [1] end-to-end system for image set selection and fusion for an all-in-focus image.

Then, classify each cell (i, j) in the grid as foreground iff the standard deviation of its focus measures across all images is above a given threshold t_1 (see Alg. 1, Line 4). Ignore background cells in Step 3.

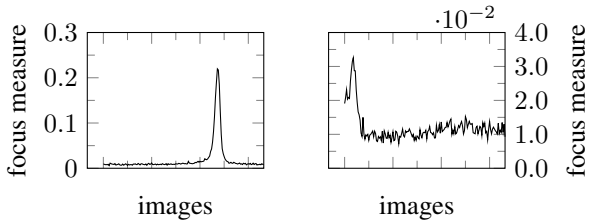

Fig. 2. Focus measures constructed for two cells from the example scene shown in Fig. 1; (left) cell with a well-defined peak; (right) cell where reliability of the peak is less clear.

Step 3. A key insight of Vaquero et al. [1] is that the problem of selecting the final set of images to fuse into a single all-in-focus image can be mapped to a set covering problem. Let A be an $m \times n$ (0-1)-matrix. A row i of A is *covered* by a column j if the corresponding matrix entry a_{ij} is equal to one. The *set covering problem* is to find a subset of the columns $C \subseteq \{1, \ldots, n\}$ that minimizes the cardinality of C such that every row is covered; i.e., for every $i \in \{1, \ldots, m\}$ there exists a $j \in C$ such that $a_{ij} = 1$. They setup a set covering instance where the rows are the foreground cells, the columns represent the images, and an entry is 1 if and only if the cell is in-focus in that image (based on whether $\phi_{i,j}(p)$ was within some threshold t_2 of the maximum value of that cell; see Lines 5–12 in Alg. 1). In general, solving set covering is NP-hard [17], however, due to the consecutive ones property present in focus stacking, selection can be computed linearly in the number of images [18].

Steps 4 & 5. Acquire high-resolution images at the specified lens positions of the set covering instance and fuse them into a single all-in-focus image.

The issue arises in Steps 2 & 3 of their pipeline, where due to hand-crafted heuristics using t_1 and t_2, some cells are incorrectly classified into foreground, background, in-focus, and out-of focus, and slightly incorrect peaks are found. We now describe improvements to address these problems.

Algorithm 1: Vaquero et al. [1] hand-crafted heuristic.

input : Focus measure $\phi_{i,j}(p)$, for each cell (i, j) and $p = 0, \ldots, n-1$; thresholds t_1 and t_2
output: Set covering instance as an $m \times n$ (0,1)-matrix $A = [a_{k,p}]$

1 $k \leftarrow 0$;
2 **foreach** cell (i, j) **do**
3 $\quad \sigma = \text{std}\{\phi_{i,j}(p) \mid p = 0, \ldots, n-1\}$;
4 $\quad$ **if** $\sigma > t_1$ **then**
5 $\quad\quad a_{k,p} \leftarrow 0, p = 0, \ldots, n-1$;
6 $\quad\quad M = \underset{p=0,\ldots,n-1}{\text{argmax}} \{\phi_{i,j}(p)\}$;
7 $\quad\quad a_{k,M} \leftarrow 1$;
8 $\quad\quad p \leftarrow M - 1$;
9 $\quad\quad$ **while** $\phi_{i,j}(M) - \phi_{i,j}(p) < t_2$ and $a_{k,p+1} = 1$ **do** $a_{k,p} \leftarrow 1; p \leftarrow p - 1$; ;
10 $\quad\quad p \leftarrow M + 1$;
11 $\quad\quad$ **while** $\phi_{i,j}(M) - \phi_{i,j}(p) < t_2$ and $a_{k,p-1} = 1$ **do** $a_{k,p} \leftarrow 1; p \leftarrow p + 1$; ;
12 $\quad\quad k \leftarrow k + 1$;

2.1. Constructing an explicit depth map

Rather than classify a cell as foreground or background (Step 2, Line 4 in Alg. 1) we construct an explicit depth map using shape from focus techniques [2, 10, 11] and use supervised machine learning to construct a classifier that predicts whether a depth estimation is reliable or unreliable. An example of an unreliable depth estimate is a plain white wall that lacks contrast or texture. Note that Vaquero et al. [1] implicitly construct a depth map and classify estimates (into what they call foreground and background) by using the standard deviation of focus measures for a cell.

To construct a depth map, we used the standard method where the lens position of the focus measure peak in a cell across all images is the estimate of the depth of the scene at that cell. The maps were improved by smoothing the focus measures for a cell to reduce depth estimate noise, and finding the peak of the smoothed focus measures. Smoothing

consisted of summing the measures of the cell under consideration and eight adjacent cells (or fewer at boundaries).

We constructed the classifier for depth estimation reliability by training a decision tree [19] based on 60 features of each cell, one of which was the standard deviation used by Vaquero et al. [1], to create a more robust classification.

2.2. Classifying in-focus and out-of-focus

Once the depth estimate for a cell has been computed and classified as reliable, the next step is to determine which lens positions around the peak are in acceptable focus. Vaquero et al. [1] used a simple heuristic where consecutive lens positions whose focus measure was within some tolerance t_2 of the peak were deemed to be in acceptable focus (Step 3, Lines 9 & 11 in Alg. 1). However, while intuitive, this heuristic relies on two assumptions that do not hold in general.

First, the heuristic assumes that a focus measure close to the peak in absolute terms is also in acceptable focus. However, focus curves often have distinguished peaks but small absolute heights. In these cases, reasonable tolerance values inaccurately deem most or all of the lens positions as in-focus. Second, the heuristic assumes that the aperture at which low-resolution images are acquired from the live preview stream is the same as the aperture at which high-resolution images will be acquired. However, to improve the accuracy of focusing and to maintain a fast shutter speed (approximately twice the video frame rate), in live preview mode the camera opens the aperture as wide as possible given the brightness of the scene. Typically, this can be as different as a wide aperture of $f/1.4$ versus a narrower aperture of $f/8.0$, and any depth of field estimate from the wide aperture would not be accurate for the narrower aperture.

Rather than estimate depth of field from the focus measures we propose to instead use standard depth of field equations to predict in-focus and out-of-focus,

$$h = \frac{f^2}{ac} + f, \qquad d_{near} = \frac{d(h-f)}{h+d-2f}, \qquad d_{far} = \frac{d(h-f)}{h-d},$$

where a is the aperture, c is the circle of confusion, d is the distance to the subject, f is the focal length of the lens, h is the hyperfocal distance, and all calculations are in millimeters. The circle of confusion is the diameter of the largest blur spot indistinguishable from a focused point source of light [20], and has established values for most contexts.

After computing a distance interval $[d_{near}, d_{far}]$, representing a depth of field in millimeters, that would lead to in-focus objects, one must compute an acceptable lens interval, representing starting and ending lens positions. This is accurately done by determining which lens positions correspond to the distance markings on a lens, and interpolating between these lens positions to determine the remaining ones. Interpolation uses the difference in reciprocals of the distances times the proportion of the distance between lens positions.

2.3. Robust selection of images via set covering

To improve image selection (Step 3) robustness, we use the explicitly constructed depth map to augment the set covering instance as follows. Let L be the set of peaks classified as reliable in the depth map. For each consecutive sequence of lens positions $p_i, p_{i+1}, \ldots, p_{i+k}$ in L, such that p_{i-1} and p_{i+k+1} do not occur in L, the lens positions p_{i-1} and p_{i+k+1} are added to L. The augmented set L is then used to construct the set covering instance, where each element of L is a row in the set covering matrix. This augmentation smooths the discrete nature of dividing an image into a grid, which addresses problems like dramatic depth changes between adjacent cells of continuous objects (such as a book angled sharply away). Smoothing makes image selection more accurate while occasionally modestly increasing the number of images selected.

3. EXPERIMENTAL EVALUATION

In this section, we perform a comparative evaluation of our improvements with the baseline Vaquero et al. [1] approach[1].

3.1. Experimental methodology

Image sets. We acquired eight benchmark image sets using a camera remote control application we implemented. A Canon EOS 550D/Rebel T2i camera was tethered to a computer via a USB cable and controlled by software, which makes use of the Canon SDK (Version 2.11). The Canon SDK does not expose functionality for sweeping the lens (Step 1, Fig. 1) so we simulated the effect by stepping the lens through the possible lens positions and acquiring a 1056×704 low resolution image from the live preview stream at each step. For evaluation, 5184×3456 high resolution images were also acquired at each lens position.

Decision tree training. We constructed labeled training data by consensus for the depth estimation reliability decision tree classifier by overlaying a grid on the low resolution images and visually inspecting each cell to determine the peak focus position (or no valid peak if the cell lacked contrast or had multiple peaks, as would occur with a blank wall or multiple occluding objects). The decision tree itself considered about 60 features based on properties of the focus measure curves and depth maps. One feature—*kurtosis*, a statistical measure based on the fourth moment of the focus measures—was the most predictive feature by far, clearly dominating *standard deviation* used by Vaquero et al.'s [1] approach. To learn the decision tree, we used Weka's J48 [21]. We experimented with parameters that led to complex trees, but reasonable settings led to trees with a single node: kurtosis. For the experiments reported here we favored simplicity at the expense of some accuracy.

[1]The implementation and data are available at: `https://cs.uwaterloo.ca/~vanbeek`.

Table 1. Number of images selected (m) and accuracy (acc.) of our method and Vaquero et al.'s [1] method compared to the minimum possible number of images needed (gold), for various benchmarks, grid sizes, and apertures. The coins and flowers benchmarks were acquired with a 200mm lens; the remaining benchmarks were acquired with a 50mm lens.

		wide		Our		Vaquero		narrow		Our		Vaquero	
grid size	benchmark	aper.	gold	m	acc.	m	acc.	aper.	gold	m	acc.	m	acc.
16×24	backyard	$f/1.4$	31	30	97.5	21	80.9	$f/8.0$	5	5	99.2	11	100.0
	bars	$f/1.4$	12	13	98.5	8	80.0	$f/8.0$	2	3	100.0	4	88.3
	books	$f/1.4$	72	81	96.0	34	55.7	$f/8.0$	11	11	99.0	16	89.8
	building	$f/2.0$	14	14	100.0	16	100.0	$f/8.0$	2	2	100.0	7	100.0
	cans	$f/1.4$	19	34	97.1	5	16.9	$f/8.0$	5	5	99.5	4	67.7
	coins	$f/2.8$	16	18	100.0	16	93.6	$f/8.0$	16	18	100.0	9	59.2
	flowers	$f/2.8$	30	33	71.7	22	38.6	$f/8.0$	14	16	89.8	10	48.6
	trail	$f/4.0$	5	5	100.0	10	100.0	$f/8.0$	3	3	99.1	4	100.0
	average		24.9	28.5	95.1	16.5	70.7		7.4	7.9	98.0	8.1	81.7
32×48	backyard	$f/1.4$	33	35	100.0	37	99.7	$f/8.0$	5	5	100.0	31	100.0
	bars	$f/1.4$	12	18	100.0	44	100.0	$f/8.0$	2	3	100.0	20	100.0
	books	$f/1.4$	87	95	100.0	76	90.8	$f/8.0$	11	12	100.0	41	100.0
	building	$f/2.0$	15	15	100.0	16	100.0	$f/8.0$	3	3	100.0	16	100.0
	cans	$f/1.4$	18	34	100.0	84	100.0	$f/8.0$	4	5	100.0	9	100.0
	coins	$f/2.8$	16	18	100.0	26	100.0	$f/8.0$	16	18	100.0	18	100.0
	flowers	$f/2.8$	35	41	91.2	163	79.3	$f/8.0$	15	19	96.1	96	93.4
	trail	$f/4.0$	5	6	100.0	21	100.0	$f/8.0$	3	3	99.6	21	100.0
	average		27.6	32.8	98.9	58.4	96.2		7.4	8.5	99.5	31.5	99.2

Parameter selection: t_1 and t_2. Vaquero et al.'s approach requires settings for the thresholds t_1 and t_2 (Lines 4, 9 & 11 in Alg. 1). For a fair comparison, we choose the optimal values. Threshold t_1 was set to the value that best fit *all* the above training data. Threshold t_2 was set by iteratively running our evaluation searching for the optimal accuracy or, within accuracy, the lowest number of images.

Performance evaluation. We compare the approaches using two performance measures: (i) number of images selected and (ii) accuracy as measured by the percentage of cells in a grid that are in focus. To compare against the minimal number of images needed and to determine the accuracy of the two approaches, we constructed a gold standard depth map for a scene using the set of *high* resolution images for the scene. We used an adaptation of 8-fold cross-validation to obtain reliable estimates of the performance of our approach (see [22], pp. 161-205), where for *each* of the eight benchmarks in turn, we trained on the other seven and tested on that benchmark.

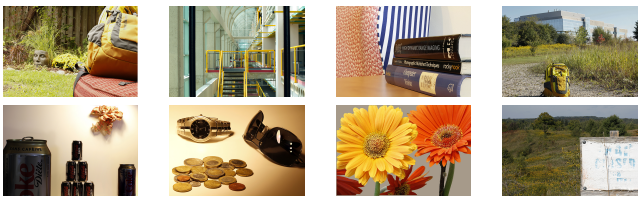

Fig. 3. All-in-focus images obtained by fusing selected high resolution images using $f/8.0$ aperture and 32×48 grid.

3.2. Experimental results

Table 1 summarizes the results of empirical evaluation. On these benchmarks, our method is more accurate than Vaquero et al.'s [1] for the coarser grid, and has comparable accuracy with many fewer images on the finer grid. For the most important case, a 32×48 grid and narrow aperture, both methods have excellent accuracy. However, our approach is close to the minimum possible number of images and a significant reduction over the number of images selected by Vaquero et al.'s [1] method with an average of 4.5 times fewer images, which would significantly reduce image acquisition time (the time between image captures often exceeds two seconds). Selection algorithm running time remained negligible. We also compared the post-processing image fusion times. In Photoshop CS5 our improvements reduced post-processing times by up to ten times. For example, for a 32×48 grid and a narrow aperture, the times (mm:ss) for fusing the images selected using our improvements ranged from 0:30 to 3:00 compared to 1:30 to 31:00 for the images selected by Vaquero et al.'s [1] approach.

4. CONCLUSION

We propose enhancements to a proposal by Vaquero et al. [1] that improves their image selection on cameras with variable apertures and lenses with longer focal lengths. Our approach maintains equivalent or better accuracy, while significantly reducing the cardinality of the selected set of images.

5. REFERENCES

[1] D. Vaquero, N. Gelfand, M. Tico, K. Pulli, and M. Turk, "Generalized autofocus," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, 2011.

[2] P. Grossman, "Depth from focus," *Pattern Recognition Letters*, vol. 5, pp. 63–69, 1987.

[3] J. Gulbins and R. Gulbins, *Photographic Multishot Techniques: High Dynamic Range, Super-Resolution, Extended Depth of Field, Stitching*, Rocky Nook, 2009.

[4] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion," in *Proc. of Pacific Graphics*, 2007.

[5] W. B. Seales and S. Dutta, "Everywhere-in-focus image fusion using controlable cameras," in *Proceedings of SPIE 2905, Sensor Fusion and Distributed Robotic Agents*, 1996, pp. 227–234.

[6] C. Zhang, J. W. Bastian, C. Shen, A. van den Hengel, and T. Shen, "Extended depth-of-field via focus stacking and graph cuts," in *Proceedings of the IEEE International Conference on Image Processing*, 2013, pp. 1272–1276.

[7] S. W. Hasinoff and K. N. Kutulakos, "Light-efficient photography," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2203–2214, 2011.

[8] S. W. Hasinoff, K. N. Kutulakos, F. Durand, and W. T. Freeman, "Time-constrained photography," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 333–340.

[9] K. N. Kutulakos and S. W. Hasinoff, "Focal stack photography: High-performance photography with a conventional camera," in *Proceedings of the Eleventh IAPR Conference on Machine Vision Applications*, 2009, pp. 332–337.

[10] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, pp. 824–831, 1994.

[11] S.-O. Shim and T.-S. Choi, "A novel iterative shape from focus algorithm based on combinatorial optimization," *Pattern Recognition*, vol. 43, no. 10, pp. 3338–3347, 2010.

[12] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining*, Morgan Kaufmann, 3rd edition, 2011.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data mining, Inference and Prediction*, Springer, 2nd edition, 2009.

[14] F. C. A. Groen, I. T. Young, and G. Ligthart, "A comparison of different focus functions for use in autofocus algorithms," *Cytometry*, vol. 6, pp. 81–91, 1985.

[15] M. Subbarao and J.-K. Tyan, "Selecting the optimal focus measure for autofocusing and depth-from-focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 864–870, 1998.

[16] H. Mir, P. Xu, and P. van Beek, "An extensive empirical evaluation of focus measures for digital photography," in *Proc. SPIE 9023, Digital Photography X*, 2014.

[17] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, 1979.

[18] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, Wiley, 1988.

[19] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[20] C. S. Johnson, Jr., *Science for the Curious Photographer*, A K Peters, Ltd., 2010.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, 2009.

[22] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*, Cambridge University Press, 2011.