Depth-Driven Computational Imaging: Portrait Mode and Privacy Filter Leveraging Focal Stack and LiDAR Data

Cameron Cipriano, Molly Housego

Boston University
Department of Electrical and Computer Engineering
8 Saint Mary's Street
Boston, MA 02215
www.bu.edu/ece

Apr. 04, 2022

Technical Report No. ECE-2022-9

Contents

List of Figures

1 Introduction

The field of lens-based photography has existed since the late 16th century [6] and has seen enormous technological improvements ranging from lens-filled pinholes in dark rooms to handheld/mobile cameras in today's smartphones. Consistent with this growth in camera technology, the miniaturization of all electronic and non-electronic components, and the exponential growth in computing power, photographic methods have consequently split into two categories: impulse imaging and computational imaging [7]. Impulse imaging techniques are those solely reproducible through physical camera systems by manual/automatic adjustments to shutter speed, aperture size, focus, and lens setup, such as extended depth of field (EDOF) imaging, motion deblurring, spectroscopy, light field capture, and illumination multiplexing. On the other hand, computational imaging can achieve (to varying levels of success) approximations to each aforementioned technique, with the addition of novel functionalities unachievable by conventional impulse imaging systems, such as image refocusing, perspective changes, omnidirectional field of view, recreation of scene structure through RGB+D cameras, stereo imaging, defocus, and diffusion [7].

In this exploration, we will be focusing on the novel functionalities granted by depth-based computational imaging, which has been powering modern technologies like Apple's Portrait Mode, introduced on the iPhone 7 plus in 2016, highlighting customer demand for the ability to control focus of an image based on depth of field. While the first iteration was not much different than how a traditional camera blurs the background of a particular subject utilizing a small f-stop, later developments introduced methods that involve merging two camera's images, or in the case of the TrueDepth camera on the iphone X, an orthogonal grid of infrared dots to measure approximate scene depth. Additionally, with the extreme and growing prevalence of facial recognition software employed by large corporations with the intent of hyper-targeted marketing, government agencies for criminal identification and overall security, and even residential buildings with the intent to replace physical keys, one's identity is becoming increasingly public [8]. Resulting from this, depth-based computational imaging can be used to

develop a facial recognition “jammer”, applying a form of visual distortion, such as blurring, pixelation, and feature morphing, to an individual’s face, preventing it from being detected entirely by facial recognition software.

2 Literature Review

2.1 Portrait Mode

Projecting light onto an image to create a depth map is not a new concept, Moreno and Noguer implemented a system that projected a visible dot grid onto a scene, such that when the subject was photographed the blur and dispersion of the individual dot provides depth information directly on the image. [1] A drawback to this method was that in order to obtain the desired subject, the dots needed to be removed and underlying color and brightness information was interpolated adding distortion. However this depth information allowed fairly accurate segmentation of the photo. This procedure allowed for not just refocusing of the foreground but any significant object in the scene as long as they are presented perpendicular to the camera.

Later Castellanos and Nguyen explored refocusing utilizing the microsoft kinect camera which implements an infrared light grid that would not be detectable to the human eye.[2] The use of infrared avoids needing to remove a visible grid from an image, however the camera had limitations yielding invalid depths at subject edges. To rectify this a hole filling algorithm was utilized setting intensity equal to the most intense pixel in a given window. Furthermore since the depth information is coming from a different image than the RGB information source the depth map must be scaled and aligned to map the information prior to refocusing. The edge information acquired from the depth map is very important when generating focus boundaries therefore hole filling and alignment procedure accuracies are essential. Castellanos and Ngueyn also utilize a machine learning alpha matte to determine if part of a scene is in the foreground or background of a focused object and improves boundary accuracy when applying artificial blurring. The limitations of this method mainly exist within the limitations of the equipment used to capture the images. The IR of the microsoft kinect is too weak to be applicable in scenes with heavy sunlight. Additionally it could not capture a high enough resolution RGB image, the depth and color information had to be acquired at separate times- meaning this application could not be used for moving scenes.

Finally there has been investigation into generating depth maps purely using information from the focal plane of an image. In theory a rapid succession of images taken at increasing focal depth will provide depth information based on blur qualities of each image in the stack. Suwajanakorn at al successfully utilized a common mobile phone, accounting for movement between images in order to acquire a depth map and refocus the image. [4] Later Wiberg was able to replicate the approach with a simpler process that consisted of alignment, generating an all focused image and a subsequent depth map with linear interpolation. [5] As computational tools improve, this method of depth perception is promising as it relies on reproducible and easily acquired equipment for image capturing when compared to light projection previously mentioned.

2.2 Privacy Filter

In order to develop a "privacy filter"/facial recognition jammer, it is important to understand the different methods for facial detection and recognition. Successful face detection algorithms of the past are heavily based on the geometrical relationships between facial landmarks, using these as the primary method of extracting facial features. As mentioned in [9], the success of these methods are highly dependent on the detection and location of these landmarks in an image, making software quite unstable in the presence of pose variation and illumination disparities. Moving from the pure geometric approach, facial recognition software began treating faces as patterns for which a more general one could find faces in an image, becoming known as the photometric approach [9]. The most successful algorithms/methods to detecting faces in images under the geometric and photometric assumptions were Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Elastic Bunch Graph Matching (EBGM). While these methods are still effective, developments in deep neural networks have demonstrated a major breakthrough in facial recognition and many vision-based tasks through the use of convolutional neural networks (CNNs) [10]. From these methods, it remains clear that these methods heavily rely on facial features and landmarks to first, detect a face, and second, recognize who the person actually is. Based on this fact, a promising method of facial recognition software deterrence is facial blurring, effectively removing all geometric and photometric information these pieces of software will typically use to identify someone. In the event blurring or major distortion is not ideal, facial morphing poses another significant viable

option as a privacy filter as it introduces high correspondences between more than one person's face, causing a purposeful non/misidentification to protect one's own identity.

Facial blurring, pixelation, and other destructive methods of concealing one's identity have been implemented before as seen on television and in research done by [11], however, visually non-destructive methods of protecting one's identity is valuable in many cases of keeping images intact while protecting one's identity.

3 Problem Statement

3.1 Portrait Mode

Many Smartphones have introduced portrait mode, designed to take a high quality image focused on a single subject. Using depth information the application of portrait mode can be expanded to multiple subjects at various distances. The following approach utilizes the focal stack of a scene in order to obtain depth information and accomplish multi subject portrait mode.

3.2 Privacy Filter

With the extreme and growing prevalence of facial recognition software employed by large corporations with the intent of hyper-targeted marketing, government agencies for criminal identification and overall security, and even residential buildings with the intent to replace physical keys, one's identity is becoming increasingly public [8]. Resulting from this, depth-based computational imaging can be used to develop a facial recognition "jammer", applying a form of visual distortion, such as blurring, pixelation, and feature morphing, to an individual's face, preventing it from being detected entirely by facial recognition software. In this implementation, fusing facial-detection bounding boxes with depth data captured via LiDAR in aligned images will facilitate facial blurring of solely facial pixels, leaving non-facial features untouched.

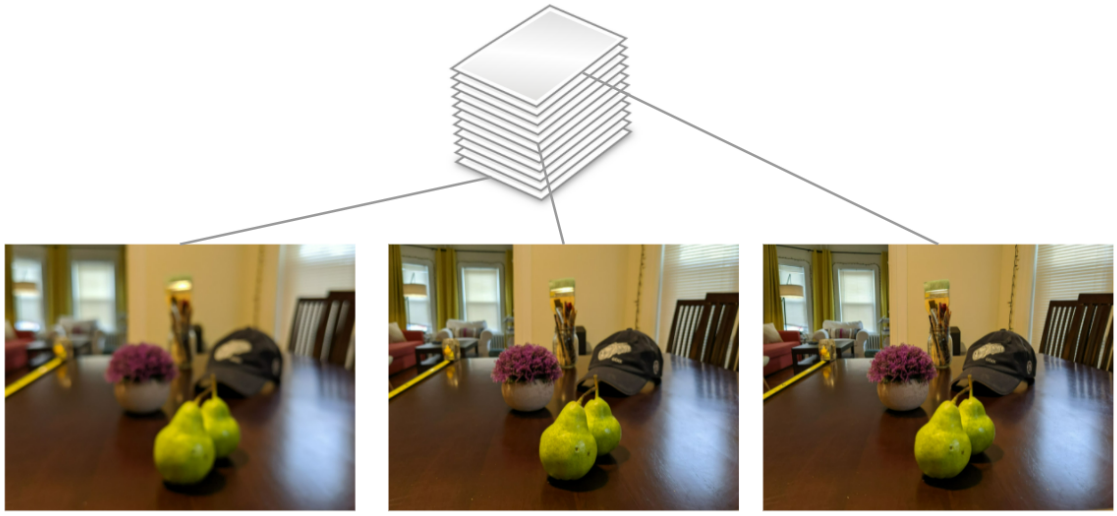
4 Implementation

4.1 Portrait Mode

In order to gather optimal depth information from blur that results from varying focal lengths of a camera lense, objects intended to be the subject of the image were positioned on the

same horizontal plane at varying distances, 13 to 50 inches away from the camera. After testing a variety of capturing methods including a DSLR, (manually adjusting focus and taking picture with remote) a smart phone, (manually adjusting focus and taking picture with remote), the method determined to be the quickest and requiring least amount of interference with the camera was using a smartphone via the “Open Camera” Application. The app has a setting called “Focus Bracketing” where the user is able to enter a distance range and desired number of images and the app will iterate through taking an image at each focal length automatically, essentially creating the unmerged focal stack that will be processed for depth information. The final setup of the portrait mode scene is seen in figure 4.1.1. Two data sets of the scene were acquired using this method: one stack of 20 images with focal length ranging from 0.1 m to 1.3 m, and another stack of 40 images with the same focal length range

Figure 4.1.1. Portrait Mode Scene Setup



Stack Image 0 Stack Image 7 Stack image 20

Figure 4.1.2. Focal Stack Image results

For computational time the 20 image stack was used to test preliminary portrait mode methods, additionally each image was resized by about a factor of 4 which added some distortion to the image but was necessary for realistic computation. The first image in the stack was taken at the shortest focal length and therefore is the image with the maximum amount of blur on each subject; this was used as a baseline for comparison to all other images in the stack. A gaussian filter was also used to create a baseline image with uniform blur however this was discarded since no image had uniform blur initially, and still resulted in bias towards whatever focal length the filter was applied to. After importing all the images, resizing and transforming to black and white double matrices. The delta between the baseline and the last image in the stack was taken, since the setup was designed such that all other factors are consistent this delta captures magnitude of blur. The delta matrix then underwent a 5 pixel range sweeping standard deviation resulting in an image that accurately reflects areas impacted by focal length change. The resulting standard deviation image between baseline and stack image 20 is shown in figure 4.1.3.

Figure 4.1.3. Stack Image 20 Blur Standard Deviation

Finally a fast forward transform was performed on the standard deviation matrix, and then filtered onto the stack image being processed. The resulting image in theory serves as a map highlighting the regions most in focus at that focal range. The map for stack image 20 is shown in figure 4.

Figure 4.1.4. FFT Filtered Stack Image 20 “Focus Map”

Once this method was defined, the process was repeated for each image in the stack. Pixel locations for each stack image called out by its respective focus map were added to a final image creating an “All in Focus” image shown in the results.

4.2 Privacy Filter

Implementation of the privacy filter thus far has been limited to RGB+D data acquisition and facial recognition bounding box generation through the use of an Apple iPad Pro 2nd generation and employment of the Haar Cascade Facial Detection Classifier algorithm. During the data acquisition process, it is essential that the device remain in the same location and the scene remain static to reduce noise in depth capture, as well as mitigate the necessity to perform a correspondence process, requiring the computation of an affine transformation between the image and depth capture to ensure points in the point cloud correspond to locations in the image at the same coordinates. During early experiments, this constraint was upheld to the best of the researcher’s ability considering lack of proper equipment, such as a tripod for an iPad, as well as the inability to remote capture images as this is an unsupported feature for the depth capture software utilized.

Once an RGB and RGB+D image were captured, they were run through the Haar Cascade Facial Detection Classifier algorithm to find all faces present in an image, returning their respective bounding boxes. In this case, the algorithm was tuned with a minimum face size of (30 x 30) pixels, a scale factor of 1.3, meaning each iteration would reduce the size of the image by 30%, and a minimum neighbor count of 4 boxes to register a face (See Figure 5.2.2). Once these bounding boxes were returned, a preliminary test of applying a Gaussian Filter to the specified region was performed (See Figure 5.2.3) to validate the effectiveness of avoiding facial detection.

From this point forward, the remaining aspect of this project involves limiting the volume of the generated point cloud to the region defined by the facial recognition bounding box, projected as a rectangular prism in three dimensions. Given this volume, the points representing the face can be segmented from the background with a method to be determined (alpha mapping, simple thresholding, etc) and reprojected to their corresponding 2D locations in the image. These pixels will then determine the face, for which blurring will be applied to.

5 Early experimental results

5.1 Portrait Mode

Figure 5.1.1 shows the black and white “All in Focus” image resulting from the methods described in 4.1 on the 20 stack dataset.

Figure 5.1.1. “All in Focus” Portrait Mode Scene

Stack Image 1 Stack Image 20 “All in Focus” Portrait Mode Scene

Figure 5.1.2. “All in Focus” Portrait Mode Scene Compared to Other Stack Images

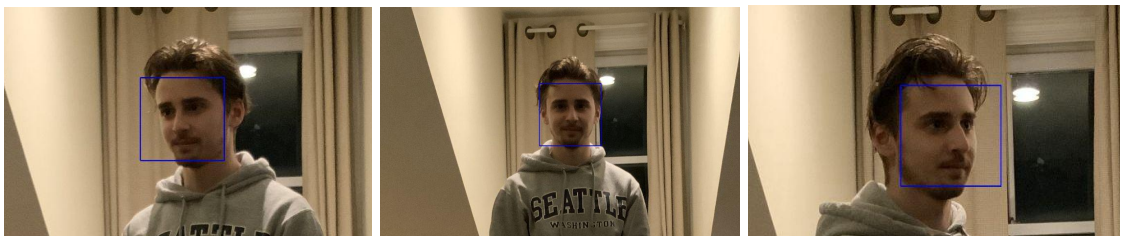
This method being a result of trial and error yielded fairly accurate but not precise results hence why the project has not progressed to quantifying depth information. When looking at the all in focus image, there is clear distortion around edges increasing in severity as the subject is further back, however to the eye this is the worst around objects not intended to be the subject of the image such as the curtains and the dining chairs. This is potentially due to very slight

misalignment during the image capturing method. There is an option to repeat the data collection with ensuring that the background is a single vertical plane perpendicular to the camera without any other features in the image, as image realignment methods are very heavy in computational costs. Additionally the focus maps of each of the 20 stack images do not capture every pixel in the image so a base using the "most in focus" image had to be used to fill in the holes.

5.2 Privacy Filter

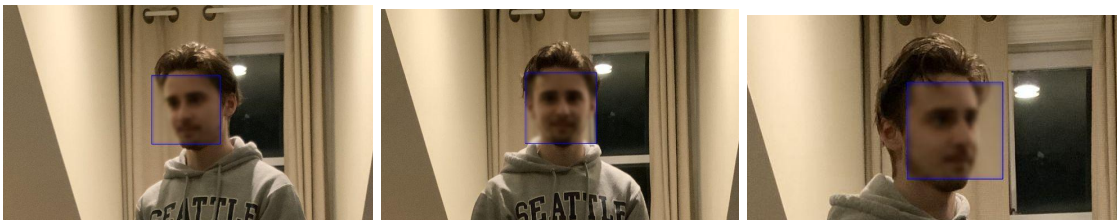
(a) (b) (c)

Figure 5.2.1: Original Photographs



(a) (b) (c)

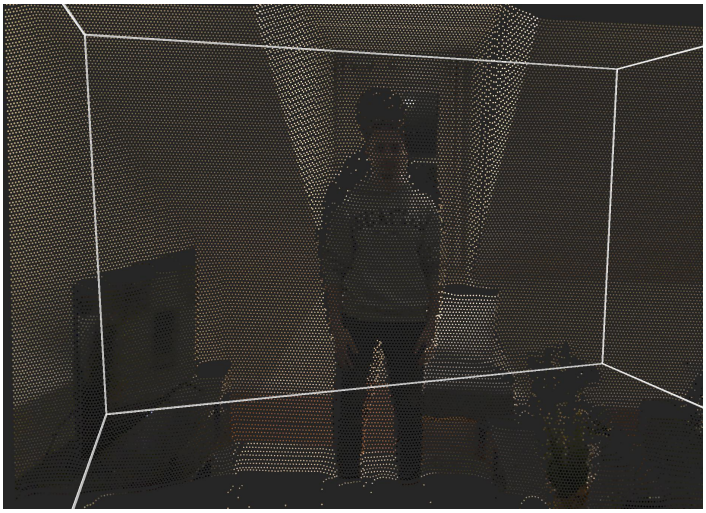
Figure 5.2.2: Haar Facial Detection Bounding Boxes



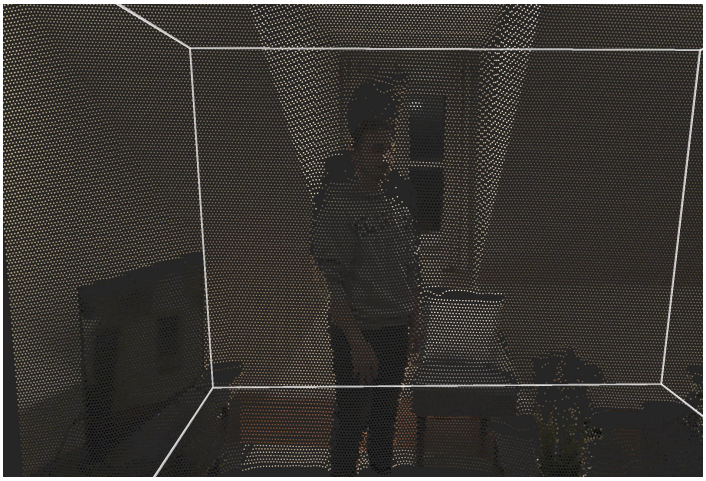
(a) (b) (c)

Figure 5.2.3: Haar Facial Detection Bounding Boxes with Gaussian Blur

(a)



(b)



(c)

Figure 5.2.4: Depth Images– (a) Corresponds to figure 5.2.1(a), (b) Corresponds to figure 5.2.1(b), (c) Corresponds to figure 5.2.1(c).

In these preliminary experimental results, it can be seen that image capture is not perfect due to the limitations of the tools available. The main disadvantage for this method as it currently stands is the separation of the traditional RGB and RGB+D images. To obtain the depth information, the application called "PointCloudScan" was utilized as it was the only one to provide a "shot" mode, acquiring 40,000 data points in a single moment, versus other applications that forced the user to capture video, amassing merely 5,000 data points per second and requiring physical motion. This requirement would have been unsuitable for this project as the denser the depth map captured, the better, as well as the desire to avoid correspondence issues.

6 Conclusions and Possible Improvements

6.1 Portrait Mode

From results in section 5.1, now that an all in focus image is captured it can be used to create a depth map of the scene, however due to the imperfections in the image it would likely be best to optimize methods up to this point. Developing an algorithm to identify the blur kernel value at each pixel would not only enable continual progress for depth information, but would potentially produce a more accurate all in focus image. The proposed plan henceforth is shown in figure 6.1.1, and comprises Molly Housego's division of labor.

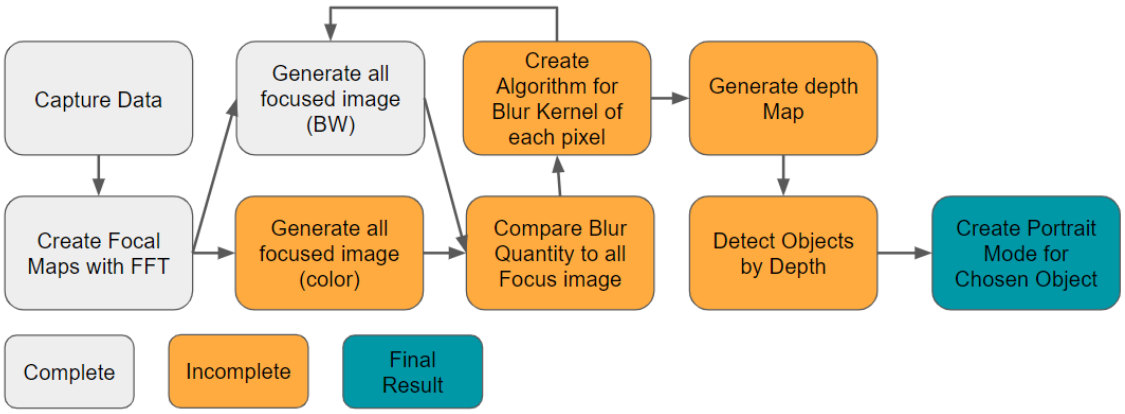


Figure 6.1.1. Proposed Project Completion Plan

6.2 Privacy Filter

The ensuing division of labor for this aspect of the project involves merging the 2D facial detection bounding box with the 3D depth data captured by the iPad Pro's LiDAR in order to

limit the region for which foreground and background segmentation needs to occur. Once this segmentation is complete, facial pixels will be identified in the depth data and need to be reprojected to their corresponding image coordinates to define facial pixels. Once these facial pixels have been identified, a blur will be applied to the face pixels alone, ensuring background pixels are not changed, and therefore implementing the privacy filter. With only the facial pixels blurred, tests of facial detection will be performed to determine the privacy filter's effectiveness for various types of blurs. As a stretch goal, multiple types of blurs/pixelation will be tested to determine the effectiveness of each against facial recognition.

References

[1] F. Moreno-Noguer, P. N. Belhumeur, and S. K. Nayar. Active refocusing of images and videos. ACM Transactions On Graphics (TOG), 26(3):67, 2007.

[2] C. Castellanos, A. Nguyen. Artificial Refocusing of High Resolution SLR Images From Microsoft Kinect Depth Maps. 2017.

[3] Y. Zheng and C. Kambhamettu. Learning based digital matting. In Computer Vision, 2009 IEEE 12th International Conference on, pages 889–896. IEEE, 2009.

[4] S. Suwajanakorn, C. Hernandez, and S. M. Seitz. Depth from focus with your mobile phone. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3497–3506, June 2015.

[5] B. Wiberg, A Method for Refocusing Photos using Depth from Defocus. Stanford University. 2018.

[6] Encyclopædia Britannica, inc. (n.d.). Perfecting the medium, c. 1900–c. 1945. Encyclopædia Britannica. Retrieved March 11, 2022, from https://www.britannica.com/technology/photography/Perfecting-the-medium-c-1900-c-1945

[7] Cossairt, O., Gupta, M., & Nayar, S. K. (2012). When does computational imaging improve performance?. IEEE transactions on image processing, 22(2), 447-458.

[8] The New York Times. (2020, July 15). Facial recognition is everywhere. here's what we can do about it. The New York Times. Retrieved March 11, 2022, from https://www.nytimes.com/wirecutter/blog/how-facial-recognition-works/

[9] Introna, L., & Nissenbaum, H. (2010). Facial recognition technology a survey of policy and implementation issues.

[10] Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R., & Busch, C. (2019). Face recognition systems under morphing attacks: A survey. IEEE Access, 7, 23012-23026.

[11] Pulfer, E. M. (2019). Different Approaches to Blurring Digital Images and Their Effect on Facial Detection.