Active Stereo Vision for Depth Reconstruction

George Younger
Stanford University
gyounger@stanford.edu

Jhonldrick Millares
Stanford University
eldrick@stanford.edu

Abstract

Nasiolabial cleft lip surgery is a procedure in which a surgeon reconstructs a patient's face to restore oral and nasal functionality in patients. This procedure requires annotation of points on the patient's face corresponding to where the face will be reconstructed. Through depth mappings of the patient's face on a computer, we attempt to assist surgeons in identifying these annotation points and projecting them onto the patient without topological deformity; that is, we aim to project the annotation points on the face after doing a depth mapping of the patient's face. Our aim was to do this with as little sophisticated equipment as possible; our approach therefore utilized active stereo with a consumer grade projector and webcam setup alongside structured light projections to calibrate the camera and projector with each other and then perform a depth mapping of the scene. We also aimed to utilize no deep neural networks to produce the reconstructions as quickly as possible at the risk of losing a bit of accuracy, while still aiming to maintain as much granularity and accuracy in the scans as possible.

1. Introduction

Cleft lip/palate is a common birth defect in many children that restricts their oral and nasal functionality. The generally prescribed approach to fixing the cleft is surgical. Through the operation, a surgeon will annotate several points on the patient's face where the reconstruction will take place; however, this is a time-consuming, laborious process that requires great effort on the part of the surgeon. Through previous work, we have constructed a neural network that will identify annotation points on the patient's face. We therefore aim to produce a depth mapping of the patient's face to be projected during surgery in conjunction with the identified annotation points to assist the surgeon. Previous work was able to project the annotation points onto an image of the patient's face, but the annotation points became deformed when being projected onto the patient's face in three dimensions. With accurate depth mappings from a scan of the patient's face, we can eliminate these deformities and make the annotations correspond to the contours of the patient's face.

1.1. Considered Approaches

In order to perform this procedure, we considered various approaches to scene reconstruction and depth mapping, namely stereo vision, active stereo vision, and using a RealSense3D camera, while operating within the constraint of using as little hardware as possible and no neural networks. Ultimately, we settled on using active stereo as this involved the least hardware: we would only need our computer in conjunction with one consumer grade webcam and one consumer grade projector, in contrast with stereo vision which would involve two webcams as well as a projector. We decided to use the RealSense3D

camera as an oracle as it represents state-of-the-art self-contained depth estimation as a camera.

1.2. Hardware and Scene Setup

In order to use this approach, we needed a projector and a camera, both of which were purchased on Amazon for less than $100 combined. We also purchased a mannequin head for our scene reconstruction as this most closely mirrored the depth mapping we would be performing on patients for cleft lip surgery.

1.3. Camera/Projector Calibration

In order to produce depth mappings for scene reconstructions from active stereo, we first needed to determine the intrinsic and extrinsic parameters of the camera and the projector, which we did through structured light. We printed out a checkerboard recommended by one of the GitHub libraries we referenced to calibrate the system.

2. Related Works

We referenced several outside papers as well as GitHub repositories that implemented the ideas from these papers. We will discuss the papers themselves in this section, and links to the GitHub repositories utilized for our project can be found in the README for our GitHub library at https://github.com/eldrickm/cs231a_project.

2.1. Structured Light for Stereo Vision

In order to implement structured light for stereo vision, we can use local homographies[2] to construct transformation matrices between the projector and the camera. It analyzes the positioning of corners across a checkerboard to determine how the positioning of a checkerboard is changed with respect to rotation and translation, and then further utilizes this information to provide us with the transformation matrices between the camera and the projector.

Figure 1. Camera and Projector Setup

2.2. Active Stereo Depth Mappings

Next, to perform our scanning, we utilized ray-plane triangulation[3] similar to the approach we used in problem set 2 to construct a depth map from various patterns projected onto the scene. These include various structured light patterns[1], including a single vertical bar that moves across the scene followed by a single horizontal bar that moves along the scene, sequential graycodes, binary codes, or XOR codes.

3. Approach

As mentioned above, our system involved utilizing both structured light to calibrate our system and active stereo to construct depth mappings once the system was calibrated.

3.1. Calibration

To calibrate the system, we projected increasingly fine-grain horizontal and vertical graycode patterns onto a checkerboard and took pictures of the projected patterns on the checkerboard with

the webcam. This process produced about 40 images (20 horizontal, 20 vertical) that were used to construct the intrinsic and extrinsic parameters of the camera and the projector. This proved to be by far the most difficult aspect of our project; in addition to retrofitting the code we found on the referenced GitHub repository to project and capture images based on our hardware setup, we found that the code was exceedingly particular about which images were acceptable to construct the parameters. It took several days of tinkering with various surrounding light settings (turning off all the lights in the room, adjusting the checkerboard so the black squares were not reflecting back at the camera, providing buffers between image captures to allow different patterns to be captured for each image) for us to get our initial parameter matrices. We believe a next step could be to iterate on top of the open source code provided by CV2 to make these image captures more robust; this would make it easier for future users to calibrate their systems without as much trouble as we had with the calibration system.

Even once we had a setup that provided us with calibration matrices, it was no guarantee that it would work again; from the point that we first calibrated the system, recalibrating it with the same approach worked around 50% of the time. It is worth noting, however, that the times we were able to recalibrate the system the parameter matrices were very accurate and provided good results, as will be seen in the Results section.

3.2. Active Stereo

Once we had calibrated the system, we retrofitted another referenced GitHub library to scan a scene for reconstruction. This involved scanning with a vertical bar horizontally across a screen and similarly vertically with a horizontal bar across the scene, after which the camera would take the captured images and reconstruct the scene based on the computed calibration matrices. After spending much of our time attempting to calibrate the system, we utilized the most straightforward, out-of-the-box implementation of depth reconstruction to compare with our baseline RealSense3D depth mapping. Given more time, we would have utilized more of the implementations for depth mapping provided by the GitHub library which we believe would improve both efficiency and granularity of scene reconstruction, but unfortunately due to the amount of effort spent calibrating the scene we were not able to pursue these avenues further.

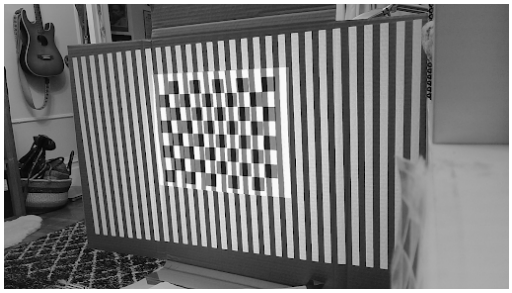

Figure 2. Binary-coded projection onto checkerboard

4. Experiments

To perform our experiments, we first set up the checkerboard on which to project structured light patterns as depicted below. This gave us our camera and projector matrices. We then adjusted the scene to put various objects in it for scanning and scene reconstruction; the scene setup, again, is depicted below. We used various objects in the scene to see how the depth mapping would adjust to some items being closer and some being farther, some being less detailed (a human forehead) and some being more detailed (a phone cord coil).

Though we tested objects of varying depth and detail, ultimately we were mainly focused on the reconstruction of the human face given its relation to the cleft lip and cleft palate surgeries discussed in the introduction, so we focused mainly on the facial reconstructions for the depth reconstructions to see how accurately we could reproduce facial features.

Figure 3. The scene setup

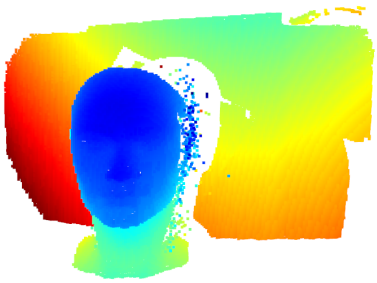

Figure 4. Our Depth Mapping

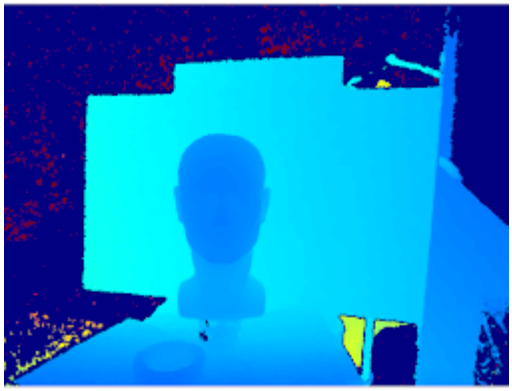

Figure 5. RealSense3D Depth Mapping

5. Results

Once we were able to calibrate the system, our results were quite successful- the depth mappings of the scene were very accurate and since we get RGB channels for free in depth mapping, we were able to reconstruct color versions of the scene to a high degree of accuracy. When comparing with the RealSense3D camera for the depth mapping, we had less granularity but the depressions, protrusions and gradients of the facial features were effectively the same- the eyes, mouth, and nose were all visible with a fair degree of detail, which was far better than what we were expecting after our initial struggles with calibrating the system.

One piece to note that our system improves over the RealSense3D camera is that it does not require a minimum depth; the RealSense requires the user to be some minimum distance away from the scene for the depth mapping before it begins to lose quality, whereas our system has no such requirements. Additionally, our system implemented the most naive scanner which gave a very rough granularity, so to have the level of detail we had even with this limitation was encouraging. Our gradient depth map measured against the RealSense3D gradient depth map is pictured above.

5.1. Limitations

One limitation of our system was that the depth from the projector's perspective can obviously not be estimated as the points were occluded. The right side of the human face is where the depth estimation starts to falter and then fall off completely into the white points pictured behind the head, which is where the camera had visibility but the projector did not. Similarly, if we were to look at the scene from the projector's perspective, we would see no depth estimation to the left of the mannequin head as those points are occluded from the camera. This is due to the translation between the camera and the projector; the RealSense camera is a self-contained unit, so there are minimal points of occlusion between the two perspectives used to estimate depth.

Additionally, because we utilized the most naive version of the scanning code to construct our depth map from the scene, our resolution was not quite up to the resolution of the RealSense3D

camera. However, as we will discuss in the future work section, there are other methods of performing the scan that will be far more efficient both time-wise and space-wise of scanning the scene, which would allow us to obtain much finer granularity scanning of the scene for the depth reconstruction.

5.2. RGB Scans

Ultimately, we felt our implementation was successful as it was extremely close to the scene we were reconstructing, the aforementioned limitations notwithstanding. Pictured below are two of the RGB scans we performed, one of just the mannequin head in front of a cardboard background and the other containing a telephone to provide more objects for the scan to analyze. As we can see, the reconstructed image looks extremely similar to the original scene, albeit less well-defined because of the process described above. We can clearly see the lips, eyes, and nose of the face are very well-defined in both of the scans shown, and we can see that the depth is preserved by the point clouds so if we were to project these scans back onto the patient's face, it would not be distorted by the contours of the patient's face like it would be if we were just to scan a 2D image.

Although as mentioned above we do lose some detail around the patient's ears, this is less important for our purposes of annotating the points where the cleft lip and cleft palate surgeries will take place because these are only on the patient's mouth and nose. If we were to perform more scans we could construct a more robust depth map of the entirety of the patient's head, but it would be unnecessary for our stated purpose.

6. Conclusion

We think through our work we implemented a system capable of producing the results we desired and with some more work we believe we could reproject the point clouds onto a patient's face to help surgeons with the annotation points.

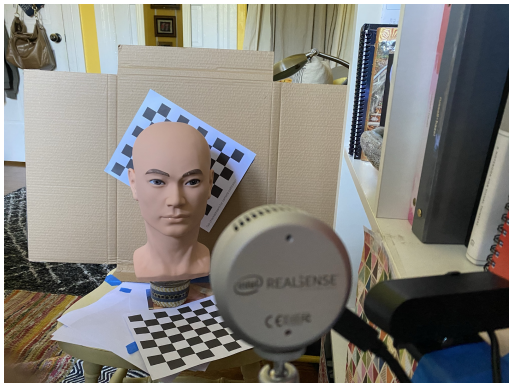

Figure 6. First Scene Setup

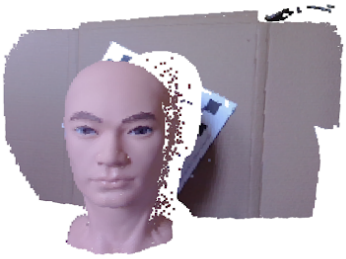

Figure 7. First Scene Scan

As mentioned above, there are a couple different avenues to move from here to improve the speed and granularity of the reconstructions. First, the naive scanner for the depth map is very slow and not spatially efficient, but for our purposes it worked well because it was easy to debug. In the future, projecting graycodes, binary codes, or XOR patterns onto the scene would be far more time and space efficient implementations as a means of getting the scene to reconstruct. The resulting spatial efficiency would allow us to store more granular images, which would lead to increased resolution in the final product- see the table included below for the differences in projected resolution (our solution would be on the order of 30 times as fine a resolution as the RealSense3D camera, which as it exists now is about 15 times as fine a resolution as our naive scanner).

Another avenue would be to use colored pat-

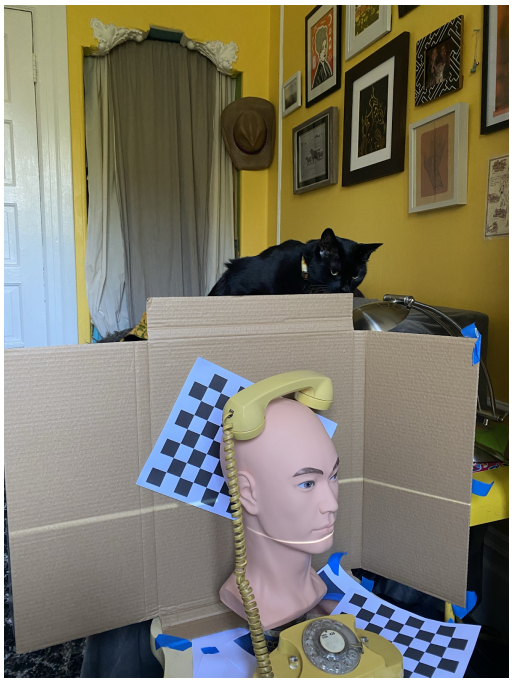

Figure 8. Second Scene Setup

Figure 9. Second Scene Scan

terns for reconstruction, as this would allow us to do one-shot reconstruction. This is advantageous for numerous reasons, chief of which would be that the patient is not necessarily stationary during the scanning. Our naive scanner requires the scene to be stationary for around 10 seconds before the scan is complete, whereas a colored pattern projection would be far quicker and would only require a single picture of the scene for the reconstruction. Additionally, this would be far

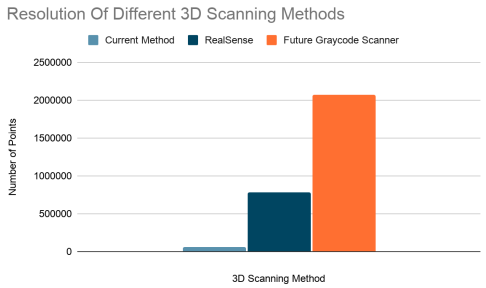

Figure 10. Second Scene Scan

more spatially efficient than the scanner implementation as it would only require a single image.

Finally, we could explore using a basic Face Detector in our scans to reject the background of the scene and focus on the face of the patient, which would lead to even greater granularity and resolution in the scan of the face. The Face Detector would give a window of reference to the scanner, which would focus accordingly on different aspects of the scene to give the best possible resulting depth map.

References

[1] J. Geng. Structured-light 3d surface imaging: a tutorial, 2011. Advances in Optics and Photonics. 2

[2] D. Moreno and G. Taubin. Simple, accurate, and robust projector-camera calibration, 2012. Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission. 2

[3] M. D. Taubin, Gabriel and D. Lanman. 3d scanning for personal 3d printing: Build your own desktop 3d scanner, 2014. Vancouver Siggraph. 2