James Roberts

Data Science and Research Portfolio

- **NBA Draft Pick Project**
    - **Poster**
    - **Research Methodology**

- **University of Texas – Applied Research Laboratories Project**
    - **Poster**
    - **Paper**

- **Stanford University – Data Science 112 Project**
    - **Poster**

- **UC Santa Barbara – Research Mentorship Program**
    - **Poster**

HAVE NBA TEAMS IMPROVED IN CHOOSING DRAFT PICKS?

Examining Selection Accuracy in the Age of Advanced Sports Analytics

1 QUESTION FORMULATION

Throughout the entire modern era of basketball which started in 1979, NBA teams have constantly sought out better ways to evaluate player performance. Driven by baseball's "Moneyball" movement, NBA teams began implementing advanced sports analytics in the early 2000s to optimize their draft selections and team rosters. **In the age of advanced sports analytics, have NBA teams improved their draft pick selection accuracy?**

2 METRIC SELECTION

All-in-One Performance Metrics Considered

BPM	RPM	PER	EPM
Box Plus/Minus	Real Plus/Minus	Player Efficiency Rating	Estimated Plus/Minus

Why BPM?

- BPM accurately represents a player's performance and is available for 1979-2019.
- BPM incorporates box score performance, team performance, and player position to rate a player's total contribution while on the court.

3 DATA FILTRATION

Steps to Ensure Impartial Player Comparison

Filtered out players with less than a full season **(82 Career Games)**

Each person represents 100 NBA drafted players (3,802 total players)

= player with 82 or more career games (1,593 total players)

= player with less than 82 career games (2,209 total players)

Eliminating players introduced unequal group sizes across draft years. Implemented Linear Regression to compensate for variability in group sizes of different draft years. Used the Pearson Correlation Coefficient (r) to objectively compare draft years.

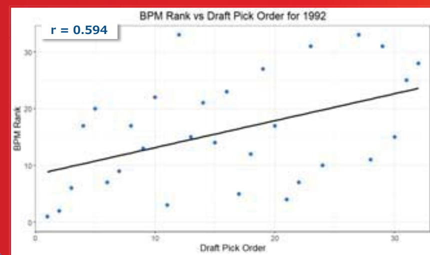
4 ACCURACY MEASUREMENT

BPM Rank versus Draft Position

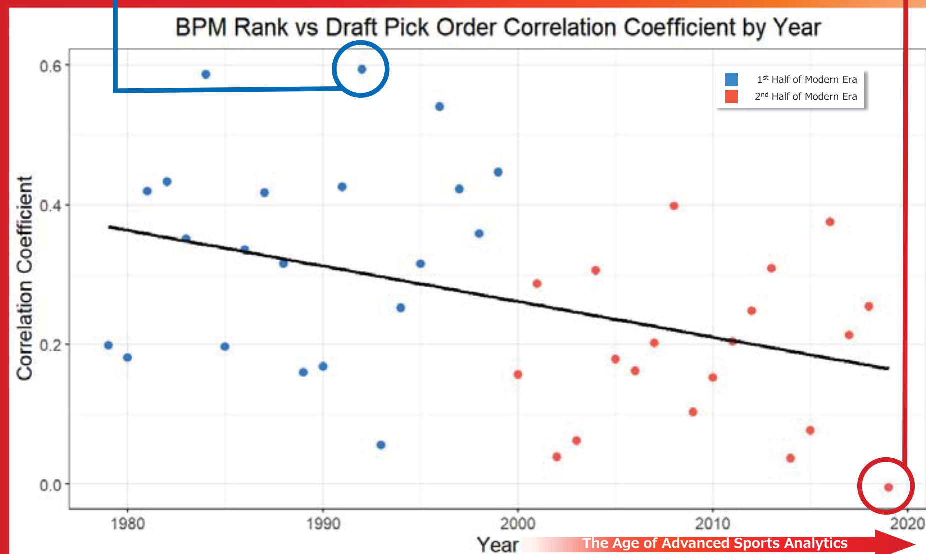
The closer the draft pick order aligns to the ranking of the players' career BPMs, the more accurate the draft year. The Correlation Coefficient of the BPM rank versus the draft pick order measures each year's accuracy.

5 ACCURACY TREND

Best Year = 1992

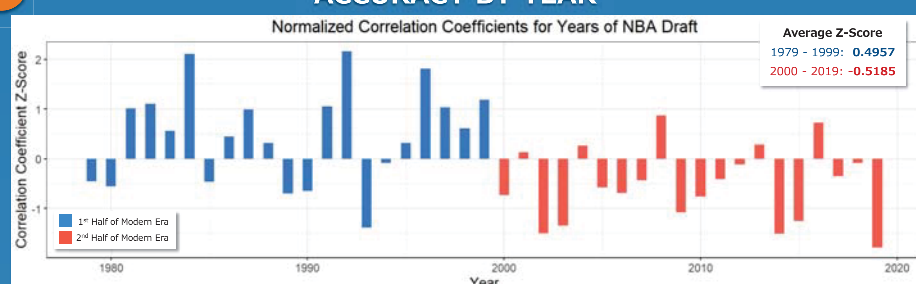


Worst Year = 2019

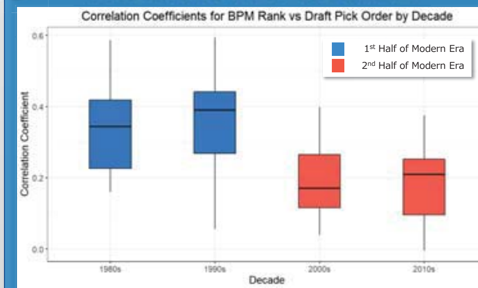


The declining correlation between BPM rank and player pick order indicates that NBA teams have gotten worse at drafting the better players earlier in the draft.

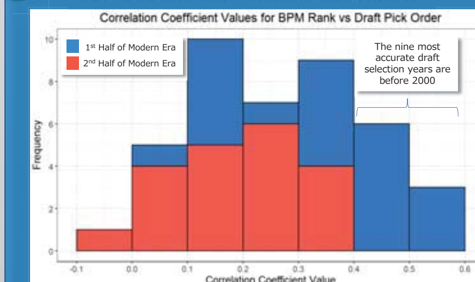
6 ACCURACY BY YEAR



7 ACCURACY BY DECADE



8 ACCURACY FREQUENCY



CONCLUSION

With the NBA's rapid adoption of advanced sports analytics in the early 2000s, the expectation is that draft pick accuracy for the second half of the modern era of basketball would improve compared to the first half.

However, **the evidence indicates that NBA teams have become worse at selecting the best players earlier in the draft since 2000.**

“Back of Poster” Details

PROJECT OVERVIEW AND QUESTION FORMULATION
While reading The Undoing Project by Michael Lewis, I was fascinated by the storyline of how National Basketball Association (NBA) teams started using advanced statistics around 2000 to better predict the performance of draft picks and potential trades. The teams wanted to take Major League Baseball's "Moneyball" techniques and apply them to basketball. It made me wonder if NBA teams had improved at selecting draft picks in the past two decades. In a perfect NBA draft, the best player would get picked first, the second-best player would get picked second, and so on. However, as we all know, this is not the reality. Every year, a few players far outperform their lower pick in the draft, while others underperform their high draft pick. But a few outliers do not tell the overall story of how accurate NBA teams are at selecting draft picks. Has draft pick selection accuracy improved during the age of advanced sports analytics?

METRIC SELECTION
The method I used to determine the accuracy of a draft year is by comparing the draft order of players to their career performance rank as measured by Box Plus/Minus (BPM). BPM is a "basketball box score-based metric that estimates a basketball player's contribution to the team when that player is on the court" (Basketball-Reference.com). I chose BPM as the advanced statistic to represent player performance because of a study by DunksAndThrees.com, a professional NBA analysis website, which compared NBA player performance metrics. Based on that study and the fact that BPM data is available for the entire period of the modern era of NBA basketball which started in 1979, it was the best all-in-one advanced performance metric for my analysis.

DATA FILTRATION
Only players who played at least 82 games, which is a full NBA season, in their professional playing careers were included in the analysis. This is because there are a few players who had amazing performances while only playing in a limited number of games, placing them above all-time great players.

PERIOD OF ANALYSIS
1979 was selected as the first year to begin the analysis as this is the beginning of the "modern era" of NBA basketball. Experts consider the early 1980s as the beginning of the "modern era" because the three-point line was added to the game in 1979 and the postseason playoffs were expanded to 16 teams in 1984. I wrangled all data required to do this analysis from the drafts of 1979 to 2019. The 2020 and 2021 drafts were not included in the analysis. 2020 was not included because its total number of players who had played 82 or more games was less than half the median number of players with 82 or more career games across all of the other draft years in the modern era. 2021 was not included because no player drafted that year had played 82 career games.

PEARSON CORRELATION COEFFICIENT (by year and decade)
Linear regression was used to account for a discrepancy in the number of players per year due to the data filtration process. By using the Pearson correlation coefficient, it was possible to compare different draft years that have different numbers of players.

ACCURACY MEASUREMENT

The closer the draft pick order aligns to the ranking of the players' career BPMs, the more accurate the draft year. The Correlation Coefficient of the BPM rank versus the draft pick order measures each year's accuracy.

Z-SCORE

The Z-score was used to normalize the data by calculating the distance in terms of standard deviations from the mean. This allows the annual draft year data to be easily compared and shows another graphical view of the data.

DATA SOURCES

Data was collected from Basketball-Reference.com and NBA.com.
Research was incorporated from DunksAndThrees.com.
Player data was collected from the beginning of the 1979-80 season until December 28, 2021.

SAMPLE DATA

Draft_Year 1992
Draft_Pick_Order_Number 1
Draft_Pick_Round 1
Team Orlando
Player_Last_Name O'Neal
Player_First_Name Shaquille
Years_Played 19
Games_Played 1207
BPM 5.1

3D Reconstruction Using a Moving Underwater Camera

James Roberts[1], Madeline Roschmann[2], Stephen Blackstock[2]
[1]Westlake High School, Austin, TX
[2]Advanced Technology Laboratory

Overview

This project focuses on creating 3D reconstructions of an underwater tank floor

3D Reconstruction:
A virtual bathymetric map made from multiple images

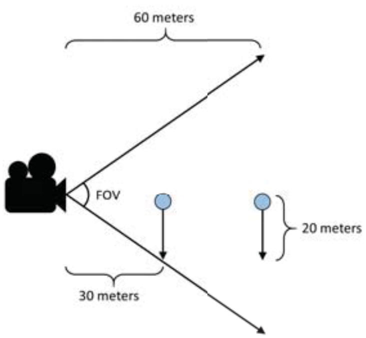
Motivation

- **Enhance** view of real-world surfaces captured
- Provide **depth**, **size**, and **identification** of nearby objects
- Assist SONAR by **verifying** detections and **differentiating** between natural and man-made objects

Panoramic of tank floor

Motion Parallax

- A camera that takes **still pictures while moving** produces images with motion parallax
- Motion parallax reveals **depth**, allowing for 3D reconstructions underwater
- Occurs due to closer objects moving a larger **percentage** of the camera's **field of view**



Demonstration of why motion parallax occurs

Objective

Obtain proof of concept for accurate 3D reconstructions of underwater scenes

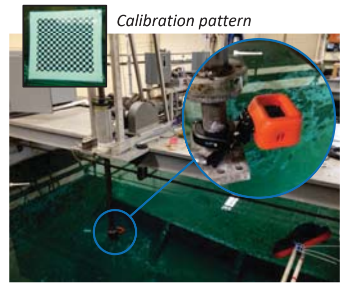
Methods

- **Camera**: GoPro HERO9 Black (Waterproof, convenient)
- **Environment**: Tank Room (Controlled setting, ideal conditions)
- **Software**: MATLAB (Toolboxes)

Data Collection

- **Calibrated** GoPro underwater to correct for lens distortion
- Mounted GoPro camera to moving platform
- Placed objects of **varying** heights to test accuracy of depth estimation
- Collected data in form of time-lapse **videos**

Calibration pattern

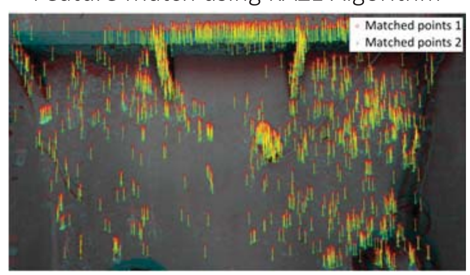


GoPro setup on platform

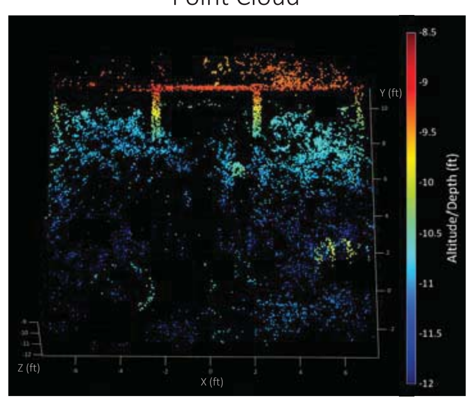
MATLAB

1. **Input** images taken of tank floor to compare
2. Used **KAZE algorithm** to detect matching points
3. Generated **point cloud** to display 3D locations of points
4. Fit surface to points for **3D reconstruction**

Feature Match using KAZE Algorithm

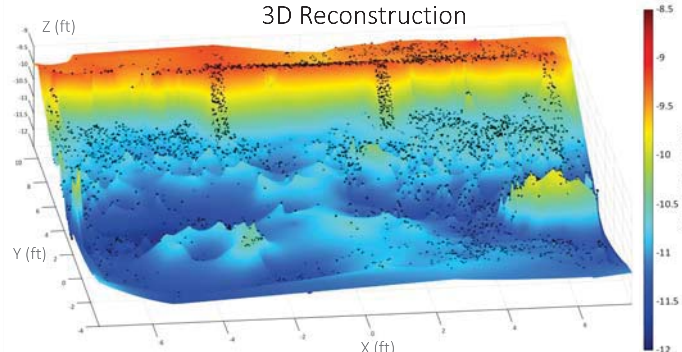


Point Cloud



Results

3D Reconstruction



Object	Measured Altitude	Mean Altitude of Points	Percent Error
Tire	11.00 ft	10.53 ft	4.27%
Barrel	10.23 ft	9.96 ft	2.64%
Santa	10.21 ft	10.24 ft	0.29%
Cliff	9.02 ft	9.28 ft	2.88%
Sand near tire	11.42 ft	11.43 ft	0.09%
Sand near cliff	10.35 ft	10.56 ft	2.03%

Conclusion

3D reconstruction of underwater scenes can be done to a high degree of accuracy in ideal conditions

Future Work

Testing thus far has been entirely within an **ideal environment**. The next step is to perform identical procedures on data from **realistic environments** that are noisy like the ocean.

Long-term goals for the project:

- Automating **object recognition** using 3D reconstructions as training data
- Building **post-processing tool** for other aquatic operations
- Decreasing **reliance** on GPS for land reconstructions

Typical ocean image

References:
https://doi.org/10.3390/jmse11050949
https://www.doc.ic.ac.uk/~ajd/Publications/alcantarilla_etal_eccv2012.pdf

3D reconstruction using a moving underwater camera

James Roberts
Westlake High School
Austin, Texas

Supervisors
Madeline Roschmann and Stephen Blackstock
Advanced Technology Laboratory

3D reconstruction using a moving underwater camera

James Roberts
Westlake High School, Austin, Texas

Supervisors: Madeline Roschmann and Stephen Blackstock, Advanced Technology Laboratory

3D reconstruction involves creating a virtual bathymetric map from multiple images to provide an enhanced representation of the real-world surfaces captured. The information necessary to reconstruct a landscape comes from an effect known as motion parallax, which occurs when taking still pictures through a moving camera. This results in objects closer to the camera moving further between each photo, revealing the depth of terrain when comparing two different frames. This project specifically focuses on obtaining proof of concept for accurately reconstructing underwater scenes. Underwater reconstructions would allow for a precise view of nearby objects and assist detection methods like SONAR to filter out natural objects and categorize man-made objects. We started by testing whether reconstruction was possible with a camera submerged in water, regardless of circumstance. We recorded videos of a controlled tank room floor using a GoPro camera mounted on a moving platform to obtain an ideal dataset. Using MATLAB's Image Processing and Computer Vision toolboxes, we compared various frames of the videos to construct point clouds that accurately portray the surroundings. The next step in our research is to determine the practicality of reconstructing unideal conditions like those in the oceans. These realistic reconstructions would be useful as testing data for machine learning algorithms to automate underwater object recognition and increase the confidence of object detection.

I. Introduction

A camera that takes still pictures while moving produces images that have motion parallax[1]. Motion parallax is one of the many depth cues humans use for depth perception and is effective at discerning the depth of features between pairs of images. Motion parallax occurs because closer objects move across a larger percentage of a camera's field of view than objects that are farther away. A visual representation is shown in Figure 1. Motion parallax can be exploited to reveal positional information, allowing for 3D reconstruction from multiple underwater still images.

For this project, a 3D reconstruction is a virtual bathymetric map made from multiple images, allowing for an enhanced view of the occurrences within a set of images. Beyond allowing for an improved look at surroundings, reconstructions are useful for proving depth, size, and identification of objects. While looking at an image can provide a qualitative analysis and identify an object, 3D reconstructions allow for a deeper view of what is occurring.

The primary motivation behind creating 3D reconstructions of underwater scenes is to assist existing detection methods like Sound Navigation and Ranging (SONAR) to increase the confidence of object detections. SONAR can effectively identify objects in the ocean but is not entirely accurate. Having a detailed and precise view of the surroundings to analyze would allow for the verification of object identifications. Increasing the accuracy of detections would allow for an efficient mapping of surroundings.

Another motivation behind testing underwater reconstruction is to establish that 3D reconstruction can be done without the help of global position systems (GPS). Creating 3D reconstructions and point clouds has been achieved on land using a drone and GPS to obtain imagery for environment observation quickly[2]. These processes have become increasingly efficient and are deployed in various situations across many fields of study, centered around enhancing the view of specific areas to gain quantitative data and insight.

However, far less tested is underwater reconstruction, which cannot use GPS for positional and timing information due to the fact that radio waves do not propagate far underwater[3]. There are many additional challenges to attempting to reconstruct underwater scenes. The difficulty of deploying and operating a system underwater is significantly higher than on land[4]. Data collection is challenging and oftentimes leads to dark and blurred images. Effects like sunlight scintillation can pose a challenge to even seemingly friendly and shallow environments. Due to the lack of formal testing, the goal was to test the viability of underwater 3D reconstructions, regardless of circumstance.

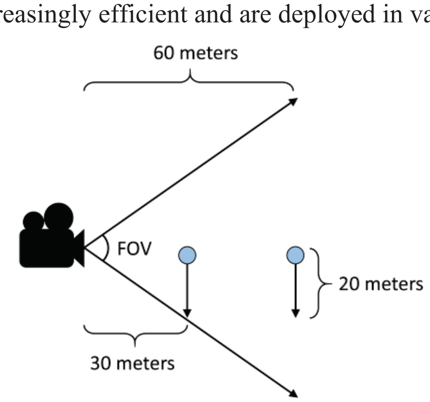


Figure 1. Demonstration of motion parallax.

II. Project Description

A. Objective

The objective of this project was to obtain proof of concept for accurately reconstructing underwater scenes. The feasibility of underwater 3D reconstruction was tested by determining the accuracy of measurements of depth, size of objects relative to tank floor, and whether the reconstructions provided the ability to identify objects in frame. If the error between the expected and observed measurements is low and the reconstructions generally reflect the surroundings, proof of concept can be established.

B. Equipment

Three main pieces of equipment had to be selected: the camera, the environment, and the software.

The camera used was the GoPro HERO9 Black, given the convenience of it being waterproof. Other cameras of consideration like Blackfly FLIR cameras required additional measures in order to record underwater, which was not necessary for establishing proof of concept. The findings within this project are applicable to other cameras in other environments,

Figure 2. Panoramic of tank room floor.

given the standardization of variables. As long as the camera parameters, field of view, resolution, and vessel speed are accounted for, the same procedures can be performed on future collected data.

The environment used was the indoor tank room within the Advanced Technology Lab, which is a controlled environment. The ability to add and remove objects and take many standardized measurements of the tank proved to be optimal for the study. Additionally, the room was ideal for detecting feature points due to the lack of additional noise, reasonable lighting, and consistent camera speed. These factors gave the best chance of establishing proof of concept for underwater 3D reconstructions. The layout of the tank floor can be seen in Figure 2.

The reconstruction software used was MATLAB, which has Image Processing and Computer Vision toolboxes that are effective for image reconstruction. MATLAB has vast documentation and high configurability, which are helpful for coding reconstructions.

III. Data Collection

Before collecting data, the conditions were established. The settings within the GoPro were held consistent throughout the study. The lens utilized was the wide lens with the hypersmooth setting off. This resulted in the air field of view (FOV) horizontally being 118 degrees, translating to 80.04 degrees underwater[5]. The image resolution was 1080 x 1920 pixels and the video format was a time-lapse with two seconds in between each frame. However, the same procedures can be applied to standard videos, as extra frames can be discarded. The movement of the platform was 0.2231 feet per second, resulting in a change of 0.4462 feet in between frames. Various objects were placed in the tank to get a better sense of the depth estimations.

The first step was to calibrate the camera underwater. Camera calibration corrects distortion by estimating the parameters of a camera lens. These parameters are also needed to perform 3D reconstruction. The estimation of radial distortion

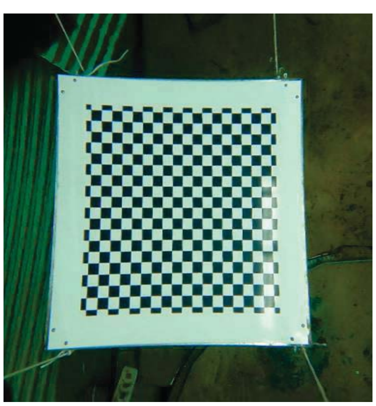


Figure 3. Calibration pattern.

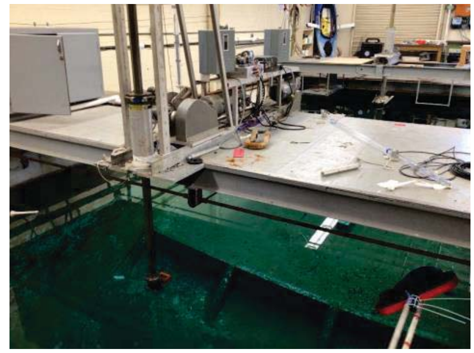


Figure 4. Testing environment.

reduces the rounding at the edges of images, while the focal length and principal point are used to calculate the relative position of the camera. The GoPro was calibrated underwater to feed realistic images into MATLAB. The calibration pattern used was a standard checkerboard arrangement, as shown in Figure 3. Any similar pattern works for calibration, however, for underwater reconstruction, the camera must be calibrated underwater as well.

The camera was then mounted onto a moving platform with a column that was able to lower the camera into the water. The platform moved horizontally at a constant speed. The camera would be lowered into the water while recording and then the platform would be moved horizontally. Refer to Figure 4 for the precise setup.

Data was collected in the form of time-lapse videos with frames spaced two seconds apart. Videos were collected of the tank by angling the GoPro to be parallel to the floor. The camera was then lowered into the water using the column attached to the platform so the lens was just below the surface of the water. The platform moved horizontally at a constant speed across the tank and then back, while the camera recorded the tank floor.

Due to the attempt to prove underwater reconstruction was possible, the conditions were made to be as ideal as possible. Ideal conditions were nearly attained since the tank was well lit, the water

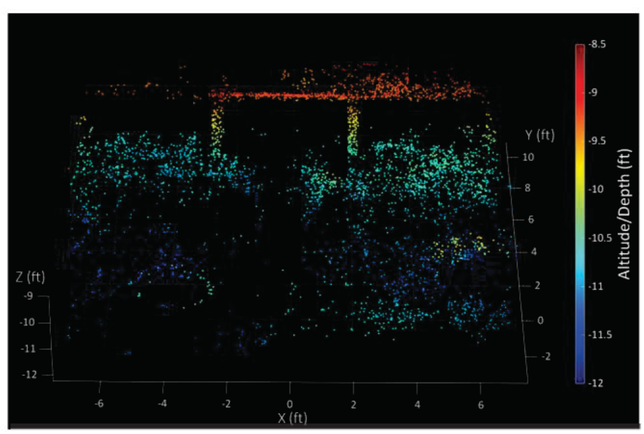


Figure 5. Point Cloud.

Table 1. Results.

Object	Measured Altitude	Mean Altitude of Points	Percent Error
Tire	11.00 ft	10.53 ft	4.27%
Barrel	10.23 ft	9.96 ft	2.64%
Santa	10.21 ft	10.24 ft	0.29%
Cliff	9.02 ft	9.28 ft	2.88%
Sand (near tire)	11.42 ft	11.43 ft	0.09%
Sand (near cliff)	10.35 ft	10.56 ft	2.03%

was clear, and there were many distinct features. This gave the highest chance for having highly detailed point clouds when analyzing the footage of the tank floor within MATLAB.

A. MATLAB

Using MATLAB's Image Processing and Computer Vision toolboxes, a set of images taken of the tank was input into the Camera Calibrator app[6]. The estimated camera parameters from calibration were then loaded into MATLAB to remove distortion in the photos.

To find matching points, the KAZE algorithm was used to detect feature points from each image and compared adjacent images to match the points[7]. The KAZE algorithm is one of the most comprehensive feature detecting algorithms, which allows for a large sample size of matched points when reconstructing. Through testing with other feature detecting algorithms like SURF and SIFT, the KAZE algorithm was determined to find the most feature points, allowing for more accurate surface reconstructions. The final reconstruction utilized 20 different images, which resulted in 19 different pairs of subsequent images being compared using the KAZE algorithm.

The feature points from the 19 pairs were used to estimate the essential matrix and determine relative motion between views. With the essential matrix, the camera pose was computed using measured values given by the camera calibration process. This allowed MATLAB to determine the real-life location of the points in relation to the camera. Then, the 3D locations of all matched points were reconstructed to generate a point cloud, which can be seen in Figure 5.

Using the point cloud, a surface was fit to estimate the depth at all locations and create a 3D reconstruction that resembled the tank floor. After creating the 3D reconstruction, other figures like a hue-saturation-value overlay on the image were built to better visualize the accuracy of the reconstructions.

IV. Results

The reconstruction is accurate based on the measurements taken of the tank floor. The objects placed in the tank are able to be identified using the

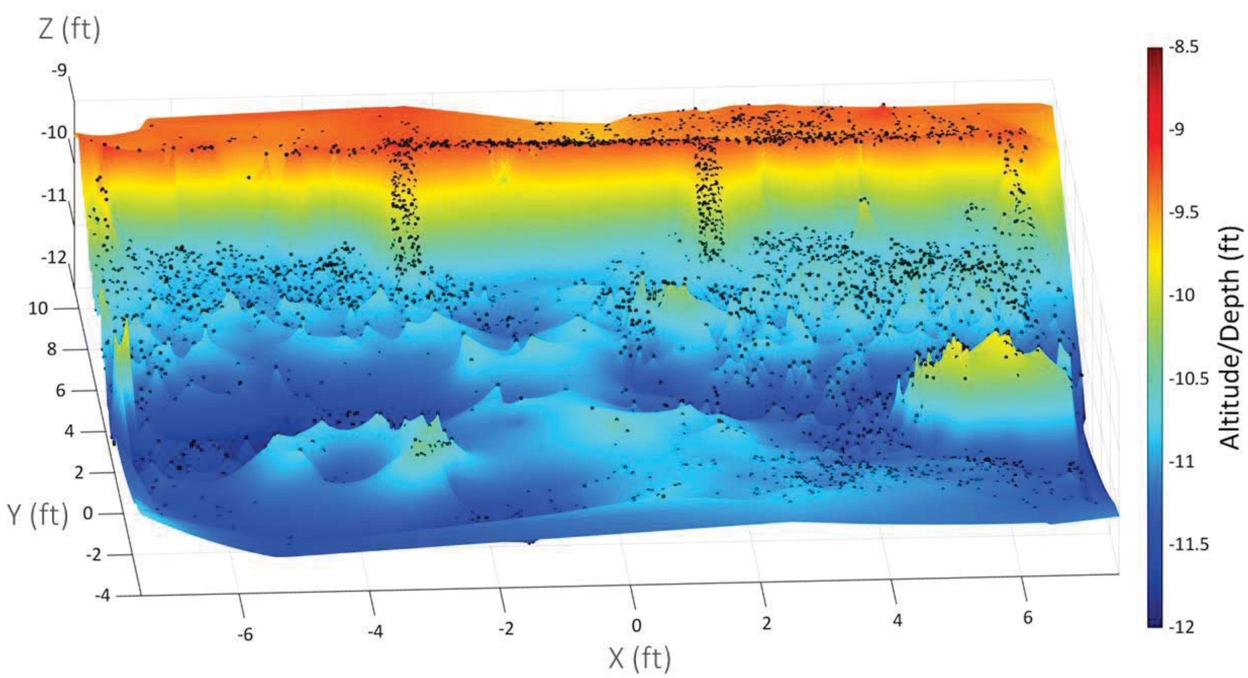


Figure 6. 3D reconstruction of surface.

reconstruction. An underwater 3D reconstruction was able to be created of the tank floor, that can be utilized for both object identification and depth measurement. The 3D reconstruction can be seen in Figure 6. The measured depth of specific points of objects within the tank room compared to the estimated depth MATLAB gave all come within 5% error, as shown in Table 1.

V. Conclusion

Underwater 3D reconstruction can be done to a high degree of accuracy in ideal conditions. This can be concluded due to the low error of depth estimation between the measured altitude and the average altitude of matched points in MATLAB and the reconstruction's general reflection of the tank floor.

VI. Future Work

While underwater reconstruction can be done to a high degree of accuracy, the work thus far has been entirely within an ideal environment involving solely linear camera motion. The next step is to perform the same procedures on data that is gathered from realistic environments that generate large amounts of noise like the ocean. Natural environments occasionally have moving objects and creatures, murkier water, and force the camera to move at variable speeds in random directions.

A. Goals

Longer term goals for the project include:

- Automating object recognition using 3D reconstructions as training data.
- Building post-processing tool for other aquatic operations.
- Decreasing reliance on GPS for land reconstructions.

Reconstructions based on data gathered within the ocean and other realistic environments could be used as data for object recognition algorithms. Currently, generating a 3D reconstruction and analyzing it can take minutes, which is impractical in possible real-time use cases. Quickly allowing for a computer to identify what is occurring within a scene would allow for more utilizations of 3D reconstruction.

A software that allows for other aquatic researchers to easily construct their own 3D reconstructions would aid marine research within all of the Applied Research Labs.

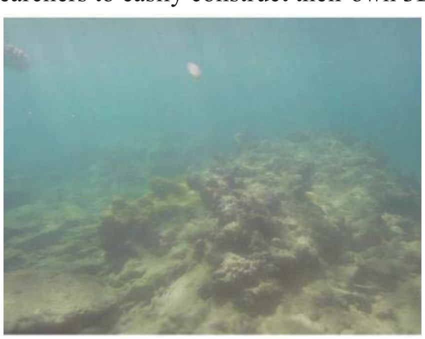


Figure 7. Typical ocean image.

While land reconstructions can be done using GPS, reducing the reliance on other technology increases the reliability of land reconstructions. While GPS may still be heavily used for land reconstructions, having the ability to reconstruct only using a drone would be useful in specific instances.

Acknowledgments

I would like to thank my supervisors Madeline Roschmann and Stephen Blackstock for allowing me to work on their project and their assistance throughout. I would also like to thank the ARL Apprentice Summer Program directors Christy Habecker and Christina Naify for giving me the opportunity to research.

References

[1] R. Brinkmann and M. Kaufmann, "Learning to See," *The Art and Science of Digital Compositing* Second Edition, pp. 15-51 (2008). https://doi.org/10.1016/B978-0-12-370638-6.X0001-6

[2] ML Cheng, M Matsuoka, W Liu, F Yamazaki, "Near-real-time gradually expanding 3D land surface reconstruction in disaster areas by sequential drone imagery," *Automation in Construction* volume 135 (2022). https://doi.org/10.1016/j.autcon.2021.104105

[3] R Meyer, "GPS Doesn't Work Underwater," https://www.theatlantic.com/technology/archive/2016/06/its-gps-underwater-for-robots/486656/ (accessed August 2, 2023)

[4] K Hu, T Wang, C Shen, C Weng, F Zhou, M Xia, L Weng, "Overview of Underwater 3D Reconstruction Technology Based on Optical Images," *Journal of Marine Science and Engineering* 11(5) 949 (2023). https://doi.org/10.3390/jmse11050949

[5] GoPro "HERO9 Black: Digital Lenses FOV Information," https://community.gopro.com/s/article/HERO9-Black-Digital-Lenses-FOV-Information?language=en_US (accessed July 30, 2023)

[6] MATLAB "Camera Calibration," https://www.mathworks.com/help/vision/camera-calibration.html (accessed July 25, 2023)

[7] PF Alcantarilla, A Bartoli, AJ Davison, "KAZE Features," https://www.doc.ic.ac.uk/~ajd/Publications/alcantarilla_etal_eccv2012.pdf (2012)

Modeling Access to City Walkability

James Roberts, Kyler Shu

Datasci 112 Winter 2024

Abstract

Walkable areas benefit citizens' physical health, boost the economy, and create a cleaner environment. Thus, walkability has become a critical factor in city planning. However, measuring the walkability of a block is a complex process that requires highly granular observations like the proximity to public transportation. Because of this, we sought to find a way to approximate the walkability of blocks based on standard metrics that are easier to collect. We also explored the relationship between specific metrics, such as income, and how they may relate to walkability. We found that walkability can, in fact, be effectively modeled with simpler metrics.

Research Question

To what extent can city walkability be understood from income and other non-spatial dependent data?

According to a 2023 survey from the National Association of Realtors, being an easy walk away from amenities is an important factor in deciding where to live for 79% of Americans. This number is skewed even higher for the younger generations.

Thus, given the high demand for walkability and the complex nature of its quantification, we aim to explore accessibility to walkable spaces and the extent to which walkability can be modeled from simple, non-spatial data.

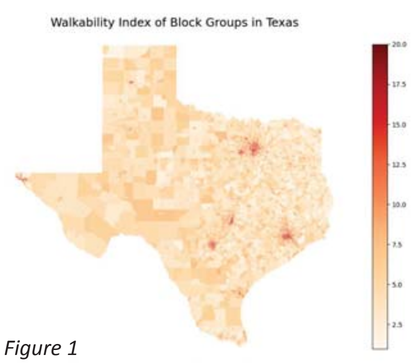


Figure 1

Data Collection

Our data collection process involved merging separate datasets and working with geospatial data.

Our first dataset, from the U.S. Government's open data website, contained data on walkability, income, and demographics but lacked geospatial information. Essentially, we had the walkability of the blocks within the United States but lacked any way to visualize the data.

To fix this, we gathered another data set from the U.S. Census Bureau containing a shapefile of the block groups within the United States. Then, we had to assemble GEOIDs from FIPS codes within our first dataset to join the two together in order to visualize our data in mapping.

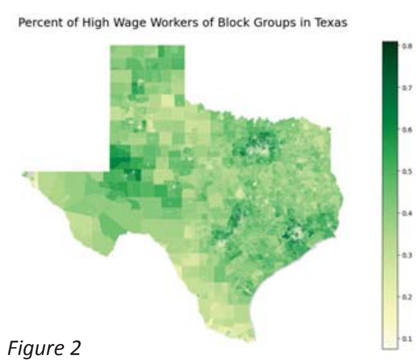


Figure 2

Data Exploration

Figure 1 maps the walkability of block groups in Texas, revealing that the most walkable areas are typically in the centers of large cities like Dallas, Houston, and El Paso.

Figure 2 maps the proportion of high-wage workers who live in a block group. This was calculated by dividing the number of high-wage workers by all of the workers in a block group. This figure reveals that the areas with the highest percentage high-wage workers tend to be suburbs, surrounding large cities. The cities tend to have a lower proportion of high-wage workers, likely because they also have a significant amount of lower-wage workers.

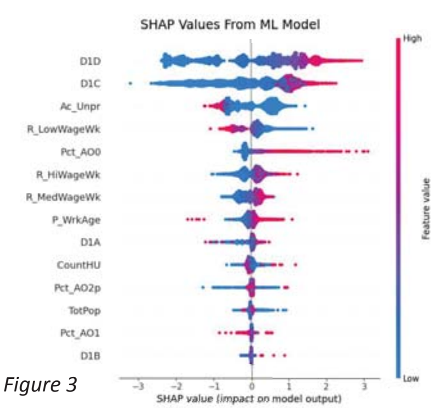


Figure 3

Data Analysis

Training a gradient-boosted regressor model on only non-spatial dependent data, we achieved a cross-validated RMSE of 2.30, with walkability scores ranging from 1 to 20.

By further applying SHapley Additive exPlanations (SHAP values), we were able to quantify which variables were most important in our model's predictions and in what direction they influenced output. Among the most impactful variables were employment density (D1C), residential density (D1D), car ownership (Pct_AO0), and income, as seen in *Figure 3*.

Conclusion

We found that walkability can be effectively modeled without data on the underlying geography of a location, greatly simplifying its quantification. Additionally, our investigation reveals a correlation between walkability and high incomes, suggesting the difficulty of accessing walkable communities. Further research could explore housing prices, and analysis could be extended to the broader United States.

Review of Recent Communication Technology Research: A Content Analysis of the Journal of Computer-Mediated Communication

UC SANTA BARBARA
Research Mentorship Program

James Roberts,[1†] America Edwards[2]
[1]Westlake High School, 4100 Westbank Dr, Austin, TX 78746
[2]Department of Communication, University of California, Santa Barbara, CA 93106
[†]corresponding author: james.robertsiii@gmail.com

Overview

- Computer-mediated communication like text messaging and social media have altered the way we interact with each other
- Researching computer-mediated communication allows us to understand the implications of media technologies through multiple lenses including education and language
- The *Journal of Computer-Mediated Communication* is an influential research journal that publishes articles covering technological communication and its importance
- While the journal has vast amounts of quality research, its overall impact has never been summarized
- Our work presents a content analysis to learn about how computer-mediated communication is being studied using SPSS Statistics
- This analysis is used to uncover past trends in communication research that can potentially identify future patterns

Computer-Mediated Communication Background

- Computer-mediated communication (CMC) is important because it has "potentially altered the way human beings share information with one another" (Liang & Walther, 2015)
- The constantly growing influence of computer-mediated communication increases the need for a content analysis covering communication journals
- The field of computer-mediated communication is understudied for how important it is for us to understand its effects on us

Journal of Computer-Mediated Communication

- The *Journal of Computer-Mediated Communication* is an open-access, peer-reviewed journal established in 1994 which has been essential in documenting the influence of technology-based communication (*About the journal*, n.d.)
- The Journal of Computer-Mediated Communication was ranked third out of 94 communication journals by the *Web of Science* for its 2-year impact factor of 7.4 (*Communication: List of journals*, 2022).
- There has been no concentrated effort to analyze the *Journal of Computer-Mediated Communication*
- Analyzing the *Journal of Computer-Mediated Communication* articles will allow for the discovery of trends in communication research like prominent theories across the journal's history

An example article

Research Goals

Increase our understanding of the field of computer-mediated communication research through a content analysis of the *Journal of Computer-Mediated Communication*

Discover patterns in computer-mediated communication like leading theories across the history of the journal

Predict future trends in computer-mediated communication technology

Determine the impact of computer-mediated communication on individuals and society as a whole

Methods

Data Collection

- The data was collected from the *Journal of Computer Mediated Communication* on July 1, 2022
- We collected 941 articles from 1995 to 2022 across 27 volumes and 119 issues
- We gathered each article's metadata from the Web of Science which includes: the author(s), their affiliations, the volume, and the number of citations
- The Web of Science was missing some metadata which had to be filled in manually
- Articles stored in Google Drive and metadata stored in Google Sheets spreadsheet
- Articles manually renamed to format that included title, volume number, and issue number

Codebook

- We used a codebook to "provide a guide for coding responses" efficiently (Carley-Baxter, 2008)
- Without a codebook, coded variables will be ambiguous, resulting in worse intercoder reliability
- We had four areas of interest:

Author Information: Author gender, location, and the number of authors

Methods: Practices taken by the article on measurements, the sample, data collection, and research setting

Theory: Which theories were employed, how many, and where they were mentioned

Technology: The device studied, application of communication, and other descriptions of the type of communication

The codebook we used

Data Coding

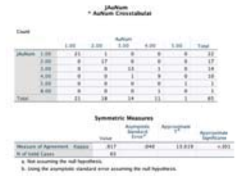
- In order to analyze our data, we needed to convert it into a format that our statistical analysis program of choice, SPSS, could interpret
- That means the article PDFs and metadata had to be coded into variables following the guidelines of our codebook
- We chose a random sample of 65 articles to code and analyze to test for reliability
- We decided on 120 different variables to code for each article in our sample to analyze in SPSS
- Each variable was clearly defined within our codebook
- For example, the variable AppComm was defined as "Is the application being studied communicative in nature? No = 0; Yes = 1" within the codebook
- We had multiple coders for each article to determine if our responses for a variable were reliable

Data Analysis

- We used SPSS Statistics to analyze our coded data
- We analyzed a sample of 65 *Journal of Computer-Mediated Communication* articles that had each variable coded by multiple people
- This sample was taken in order to establish reliability before coding the entire dataset
- We performed an interrater reliability test to check if the different coders had similar responses for each variable
- For variables of interest, we plotted bar graphs and other forms of visually displaying the data

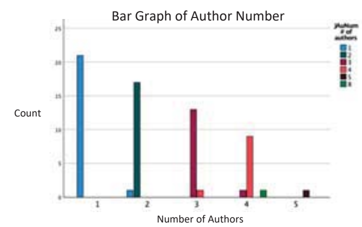
Results

- After performing an intercoder reliability test for each variable, we figured out that the reliability was lower than desired for many variables
- We were ideally looking for above 90% agreement between the coders
- An example of a variable that was reliable is AuNum, which is the number of authors who contributed to writing an article
- Across the 65 articles analyzed, our Measure of Agreement was 0.917, meaning the coders agreed 91.7% of the time



SPSS output for reliability test

- 21 articles had a single author, 18 had two authors, 14 had 3 authors, 11 had four authors, and 1 had five authors.



Discussion

- When we took a small sample of our final dataset and performed intercoder reliability tests, our agreement for many variables was too low to establish reliability
- This was likely due to a difference in understanding the methods of the research performed in the articles
- We have to do coder training to make sure that all coders are on the same page and get the same responses when coding
- We were still able to establish reliability on a portion of variables, which help us to understand the distribution of who is researching communication technology and where this research is occurring

Conclusion

- There were many findings about the *Journal of Computer-Mediated Communication* that came out of our analysis but one of interest is author location
- 64.6% of authors within our sample of 65 articles were located in North America at the time of publication
- This may be because of the large number of universities that support communication research within the United States and most research being conducted in English
- However, since the journal is an International Communication Association journal, expanding the scope of location would improve our understanding of how computer-mediated communication operates around the world

Future Work

- We plan to analyze the journal, *Computers in Human Behavior*, and compare those results to that of the *Journal of Computer-Mediated Communication*
- We will use Python, RStatistics, and Linguistic Inquiry and Word count to perform these analyses
- We will employ both human coders and machine coders to compare the accuracy and determine which is more efficient

Acknowledgements:
Joseph Walther (Director of the Center for Information Technology and Society, UCSB)
Ron Rice (Distinguished Professor, Department of Communication, UCSB)
Nicole Ellison (Editor for JCMC)
Emily Shen and Mahek Patel (Research Assistants under America Edwards)
Raphael Ruschel Dos Santos (TA)
Dr. Lina Kim (Program Director, RMP)

References:

[**1**] Liang, Y., & Walther, J. (2015). *Computer Mediated Communication*.
[**2**] *About the journal*. Academic.oup.com. (n.d.). Retrieved July 10, 2022, from https://academic.oup.com/jcmc/pages/About
[**3**] *Communication: List of journals*. OOIR. (2022). Retrieved July 24, 2022, from https://ooir.org/journals.php?category=communication
[**4**] Shamrao, Gajbe & Sharmrao, Dr.Gajbe & Shamrao, Gajbe. (2021). Content Analysis of "Journal of Health Communication". International Journal of Research in Library Science.
[**5**] Simonson, P., & Park, D. E. (2016). Introduction. In P. Simonson, and D. W. Park (Eds.), The international history of communication study (pp. 1–22). Routledge. https://ijoc.org/index.php/ijoc/article/view/9060
[**6**] Carley-Baxter, L. (2008). Codebook. In P. J. Lavrakas (Ed.), Encyclopedia of survey research methods (pp. 101-101). Sage Publications, Inc., https://dx.doi.org/10.4135/9781412963947.n69
[**7**] *Divisions and Interest Groups*. International Communication Association. (n.d.). Retrieved July 25, 2022, from https://www.icahdq.org/page/div_igs
[**8**] WanMansor W.F., Zakaria M.H. (2017) Researching Computer-Mediated Communication. In: King K., Lai YJ., May S. (eds) Research Methods in Language and Education. Encyclopedia of Language and Education (3rd ed.). Springer, Cham. https://doi.org/10.1007/978-3-319-02249-9_41