Gaze Tracking: Seeking Critical Information for Autonomous Excavation

by

Kaymie Shiozawa

Submitted to the
Department of Mechanical Engineering
in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Mechanical Engineering

at the

Massachusetts Institute of Technology

June 2019

Signature redacted

Signature of Author: ______________________________
Department of Mechanical Engineering
May 10, 2019

Signature redacted

Certified by: ______________________________
Harry H. Asada
Ford Professor of Engineering
Thesis Supervisor

Signature redacted

Accepted by: ______________________________
Maria Yang
Associate Professor of Mechanical Engineering
Undergraduate Officer

Gaze Tracking: Seeking Critical Information for Autonomous Excavation

by

Kaymie Shiozawa

Submitted to the Department of Mechanical Engineering
on May 10, 2019 in Partial Fulfillment of the
Requirements for the Degree of

Bachelor of Science in Mechanical Engineering

ABSTRACT

Automating excavation in mining and construction applications is crucial today as the supply of skilled operators cannot match market demand. To efficiently make control decisions for autonomous excavators without having to take in all visual inputs from a typical operator's field of view, gaze tracking is employed in solely extracting key visual information that skilled operators use in the field.

Both a front facing camera depicting the world view of the subject and two eye facing cameras that track the subject's pupil movement are worn by a subject to identify regions and features that are of high interest to operators during a digging task. Key features, such as the interface between the soil and the bucket, are characterized using U-Net, a Convolutional Neural Network designed for image segmentation.

Through this study, key regions, the inside of the bucket and the opening of the bucket, as well as key features, the soil-bucket interface, were identified to be of high interest to subjects. This information can serve to identify only the necessary visual inputs in the control decision process, thus shortening computation time.

Thesis Supervisor: Dr. Harry H. Asada
Tile: Ford Professor of Engineering

Acknowledgments

The author would like to thank Professor Harry Asada for providing insight into the direction of the project, Mr. Filippos Sotiropoulos for his advice in carrying out the experiments, Mr. Wenbo Tao and Mr. Mohamed Abdelhamid for their technical guidance in writing this paper, and Ms. Jessie Stickgold-Sarah and Mr. Keith Clavin for their recommendations on better communication techniques.

Table of Contents

List of Figures

1. Introduction

The operation of mining excavators is a demanding and often undesirable job that requires operators with a high skill level. Obtaining this level of skill entails years of training, and practice on lower capacity equipment. This has led to a situation where the supply of skilled operators cannot match market demand [1]. To solve the shortage of human operators, the overarching objective of this project is to develop the methods and techniques required to automate excavation in mining applications. Previous research has focused on more minor details and subtasks of autonomous excavation, such as slippage control [2] or low-level safety protocols [3]. However, to tackle the excavation operation as a whole, there are simply too many inputs to process efficiently. This project is novel because it aims to isolate key information from human operators' gaze tracking data while they complete any subtask of excavation, which can then be implemented to the controls decision making of an autonomous excavator. With the information from the gaze tracking experiments, high volume data that previous researchers were not able to tackle, can more easily be sorted to only focus on the essential qualities of an excavation process.

During a dig, an experienced operator not only looks at the bucket filling rate of the current cycle, but also takes into account the consequence of the current scooping and its impact on the succeeding scooping and the terrain profile as a whole. Furthermore, an operator scans the surrounding conditions of the work site: the excavator stability and ground conditions, the location of dump trucks, and other obstacles and significant constraints. Experienced operators are trained to sort through large amounts of visual inputs and take in important data points quickly, but quantifying all of these features in an autonomous system makes the input space dimension intolerably large. It may be infeasible to include many thousands of features and conduct experiments to obtain statistically significant data from which a practical control algorithm can be extracted. However, many of these features may not be needed to make a control decision in every time step. Rather, these features may play a critical role at particular points in a sequence of operations, or under specific conditions and contexts. Finding such critical features and determining when an experienced operator uses such features is crucial in applying the advanced skills of the operators to the automated machine.

Since this project aims to identify the key visual features that operators focus on during digging, gaze tracking is used at first to gain a better understanding of the locations that operators fixate on in their field of view. A camera that looks into both the eye and the surroundings is worn by a subject to collect data. Through statistical analysis of the gaze concentration in predefined regions, such as in or out of the bucket or quadrants of the bucket, trends in gaze tracking data is obtained. Furthermore, predetermined features, such as the interface between the bucket and the soil, is identified using Convolutional Neural Networks and later correlated to the gaze tracking data to verify that the feature is relevant to human operators. The results show that there are in fact key regions and features that operators visually focus during a dig, which can, in the future, be implemented to isolate the visual input into a controls algorithm.

Identifying key features and regions in a larger field of view can be beneficial in improving processing time especially for real-time data collection. For example, optical flow analysis, a method used to measure soil flow into the bucket, may be more computationally efficient if the analysis were to be conducted in a small portion of a bucket than the entire bucket. Moreover, through this project, this "small portion" will be determined as a "key region or feature" that is essential to an experienced human operator, making the dig an optimal operation with little inefficiency. The processing of information in autonomous excavation can be thus improved in speed, and the operation itself, such as digging, can become more efficient, through ensuring that every dig maximizes the soil volume in the bucket, for example. This method can also be applied to tasks other than digging in the field, or perhaps even other manipulation tasks, such as driving a crane.

2. Related Work

Previous research in autonomous excavation have explored roughly four levels of autonomy: teleoperation, sharing control between the human and the machine, automation of the selection of digging location, and automation of digging over a long period.

The first level, teleoperation, usually removes an operator from the excavation site, often for reasons like safety, but the controls are still done by the human operator. For example, there are projects in the field pertaining to uncovering of buried munitions [4], waste [5], or utilities. Although this method resolves the concern for the safety of human operators, it does not provide a solution to the shortage of skilled human operators in the field.

The research in the second level of automation, where the human shares the control of the excavator with a machine [6–9], often focus on the digging task. The operator decides where the dig will take place, and the excavator takes over the digging procedure using force and position feedback loops. This method provides an opportunity for workers that lack the skills to efficiently maneuver an excavator to complete the task, but does not fully remove the need for humans.

The third level of autonomy is where the machine can automatically select the desired dig position and conduct a dig. These systems use measurements of the topology of the terrain using ranging sensors to optimize the dig for maximum volume of soil per cycle [10]. To further optimize digs, Neural Networks have also been employed to model soil conditions during operation [11]. In this level of autonomy, researchers have simplified the excavation process through hand-coded scripts focusing on subcategories of excavation including, the digging task performance, soil modeling, or bucket-soil force interaction [12]. Tasks are separated into scripts with parametrized inputs from the environment for real-time planning and execution of tasks [13].

To improve upon the previous level, the highest level of autonomy currently in the field automates a sequence of digs over a long period of time [10]. However, some of these systems are yet to match the performance of expert human excavators, as they are unable to model every situation possible and cannot adapt to shifts in environments thus necessitating intervention by human operators.

To fix this, parametrized scripting, an adaptive free motion planning approach, is applied to excavation to take in information about the excavator machine's performance and output optimal parameter values. For example, for an input task called, "digging", the output may be the angles and the velocities of the bucket and the boom [14]. The success of this research can be enhanced by taking the approach of this paper, and inputting information not only about the machine but also about a human operator's decision-making process. More recent studies have utilized neural networks to mimic a human operator's closed control loop mechanism, yet there are simply too many visual data inputs to efficiently process in real-time [15].

To overcome the problem of inefficiencies in data processing due to the abundance of visual inputs, gaze tracking of a human operator is proposed in this paper to pinpoint key features and human habits. Gaze tracking technology has been used in a variety of applications including advanced driver assistance systems to understand the driver's focus points and situational awareness [16]. Moreover, some work has been done to apply deep learning to gaze tracking to determine a person's object of interest and conduct object classification [17]. Both these works utilize the information from a person's gaze to identify what is important to them and apply it to automating a system. On the other hand, the gaze patterns of a human excavator operator have yet to be analyzed. As in other fields, such as autonomous vehicles, gaze tracking provides this project a different angle to the automation of excavators, in which a human operator's attention points can be directly applied to the controls protocol.

3. Gaze Tracking Experimental Set Up and Procedure

In order to identify key visual attention points in the field of view of a human excavation operator, their gaze needs to be tracked while conducting a typical task, such as digging. Trends in the gaze tracking dataset across multiple digging trials was used to at first characterize regions of interest. Then, the gaze tracking data was visually monitored to qualitatively understand certain features that are of high interest.

The human operator subject was requested to wear a gaze tracking device and simultaneously manipulate an excavator by looking at a screen that was live streaming a video from a camera filming the excavator's movement. The goal of the subject was to complete a digging and picking up task as indicated in Figure 3-1.

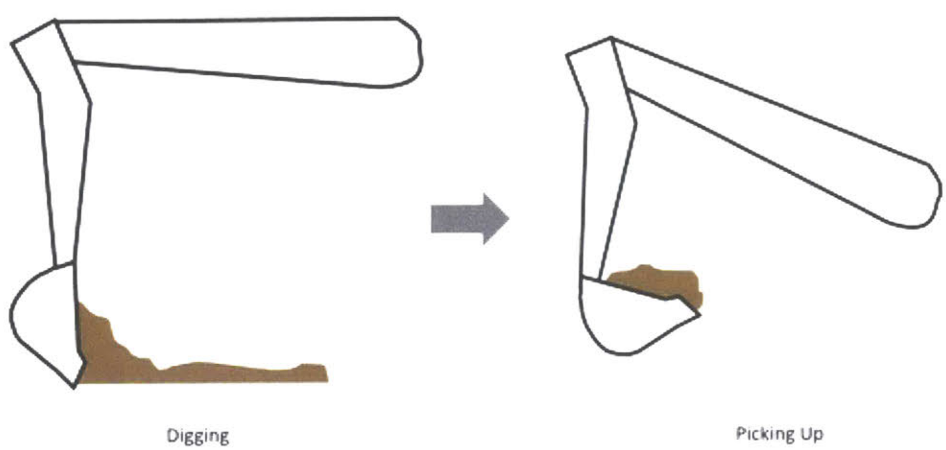

Figure 3-1: Transition from digging to picking up phase.

The gaze tracking experiment had two set ups. The first set up consisted of a larger field of view mimicking a typical excavator operator's field of view, where not only the mound of soil of interest was in sight, but other obstacles, such as other mounds of soil and rocks, was also in view. The operator was asked to manipulate the bucket in a digging motion while looking at only the field of view provided. This experiment was done ten times for one operator.

The second part of the gaze tracking experiment was conducted to further identify specific regions of interest in the field of view of only the bucket and the surrounding soil. The subject was requested to conduct a digging process similar to the first set up, but the camera was mounted on the arm of the excavator and pointed at the bucket, as shown in Figure 3-2. This way, the bucket does not translate with respect to the camera, but simply changes its angle depending on the operator's maneuver. This experiment was done 30 times each for two different operators.

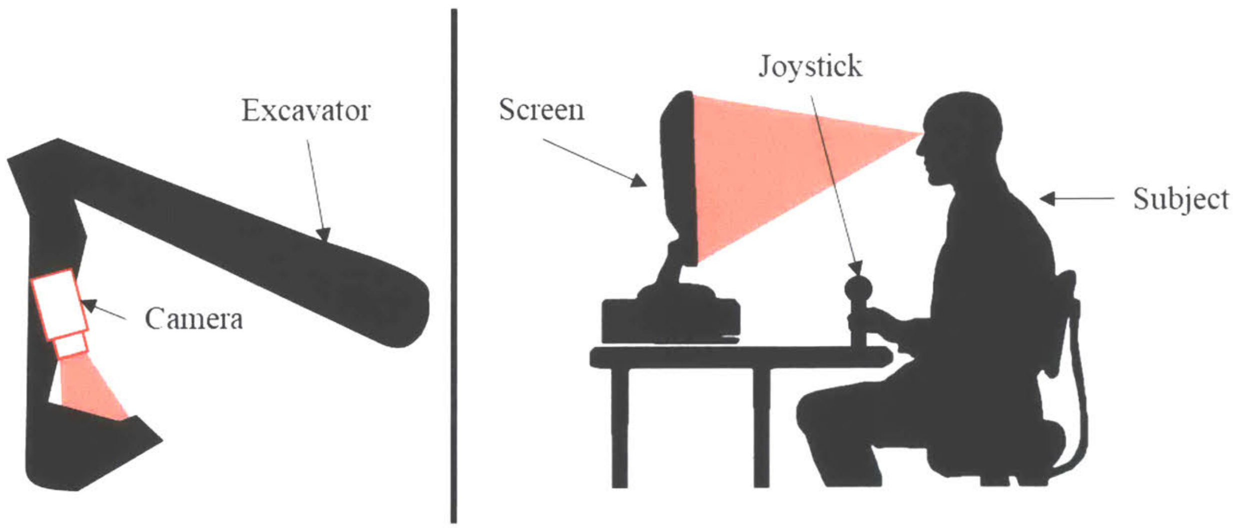

Figure 3-2: Experimental Set-Up. The subject sits in front of a screen. The screen displays the viewpoint of the camera that is attached to the arm of the excavator. The subject is able to operate the excavator using a joystick.

The excavator used in this experiment was a prototype created in the D'Arbeloff Laboratory at MIT. The prototype excavator has two degrees of freedom, where the "boom" and the "arm" is controlled at the base using a four bar linkage mechanism, and the bucket is controlled at the end of the arm. The subject can operate the prototype excavator using two joysticks: the right joystick controls the bucket and boom angle, while the left joystick controls the arm angle.

The gaze tracking device used were the Pupil Headsets, which has a front facing camera, with a sampling frequency of 120 Hz, depicting the world view of the subject and two eye facing cameras, with a sampling frequency of 200 Hz, that track the subject's pupil movement. The gaze accuracy and precision are 0.60 degrees and 0.08 degrees respectively. The distributor, Pupil Labs, provides the calibration procedure and gaze estimation features. A surface heat map, as shown in Figure 3-3, is created from the tracking data using Pupil Labs' Offline Marker Detection and Surface Heat Map Plug-In, in which the subject's focus points are quantified based on the frequency of gaze in a specific portion of the map. Furthermore, fixation detection algorithms can be added to detect the difference between a fleeting glance and an intentional gaze by using a dispersion based algorithm mentioned by Barz et al. [18].

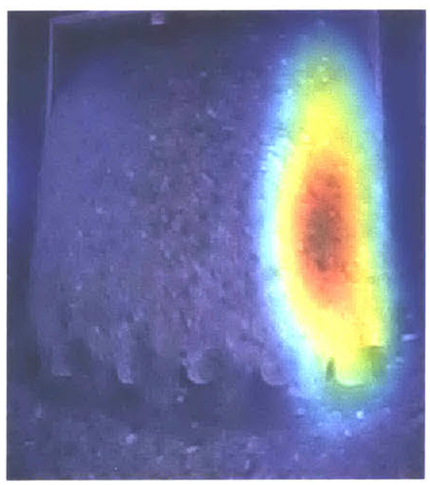

Figure 3-3: A surface heat map indicating high and low gaze concentration areas.

4. Convolutional Neural Networks to Identify Features

To quantify subjects' gaze distribution on a feature, a Convolutional Neural Network (CNN) was used. In particular, U-Net [19], a CNN suited for segmentation on fewer training images was used for this application. As shown in Figure 4-1, the network consists of a contracting and expansive path, creating a u-shaped architecture. The contracting path reduces the spatial information while increasing the feature information, and contains convolutional layers, each followed by a rectified linear unit (ReLu) and a max pooling operation. The expansive path combines the feature and spatial information through deconvolution and concatenation of the corresponding cropped feature map from the contracting path. This architecture that combines the location information with the feature information is highly beneficial in segmentation. Furthermore, because this U-Net conducts massive data augmentation, it is also suitable for this paper's application where the number of annotated samples is limited.

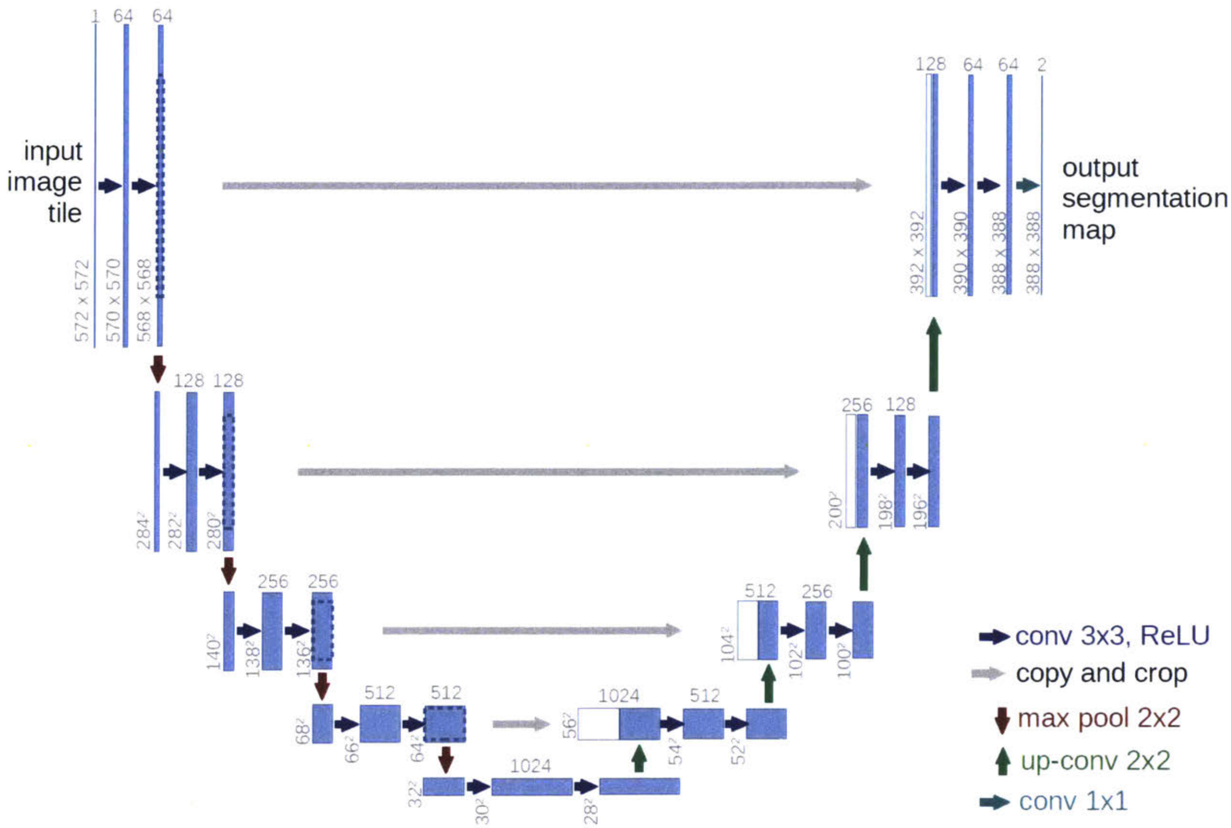

Figure 4-1: U-Net architecture. Each blue box represents a multi-channel feature map. The number above the boxes correspond to the number of channels. The x-y-size is denoted at the left edge of the boxes. White boxes are copied feature maps. The arrows represent the different operations done on each layer.

Fifty images taken during a digging operation with a camera attached to the arm of an excavator were annotated using Polygon Rnet++'s edge detection algorithm [20]. The feature of interest, the soil-bucket interface was segmented. Ten of these images each were allocated for test and validation sets. Figure 4-2 shows an example of a labeled image, indicating the soil-bucket boundary.

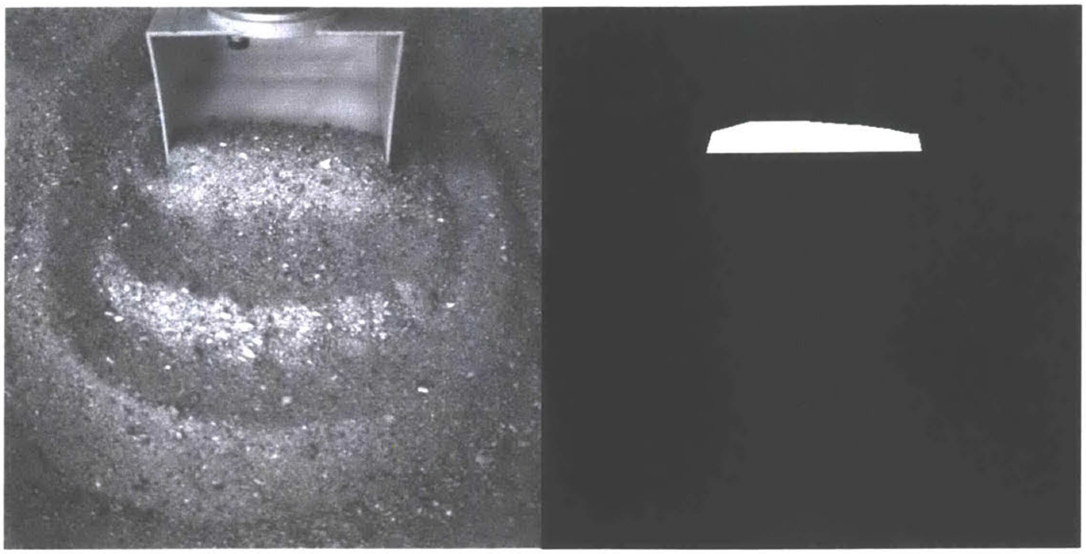

Figure 4-2: An example of an annotated image. Using Polygon Rnet++'s algorithm, the 50 images were annotated by hand to use for training. The soil-bucket boundary, and a small area of soil nearest to that boundary is selected as a "key feature".

5. Experimental Results

The trends in visual data that are identified by a human operator to be key inputs during a digging operation were measured through gaze tracking experiments and analyses. The number of fixations, intentional gazes, were measured for every digging trial and plotted with respect to predetermined regions, such as quadrants of the bucket. Further analysis of gaze tracking data was done through investigation into the correlation of gaze locations and features, such as the transition point between the soil and the bucket, instead of predetermined regions. Convolutional Neural Networks were used to identify features.

5.1 Gaze Tracking – Identifying the Bucket as a Key Region

In the first gaze tracking experiment set up, the subject's field of view contained not only the bucket, but also the surrounding soil and rocks. This experiment displays the operators' interest in the bucket region for an average of 98.2 ± 0.3 % (95% confidence interval (C.I.)) of fixations tracked across all trials. Figure 5-1 shows the distribution of gaze in and out of the bucket region. This consistent interest in the bucket region is visually depicted in Figure 5-2 where the gaze distribution graph is red over the bucket.

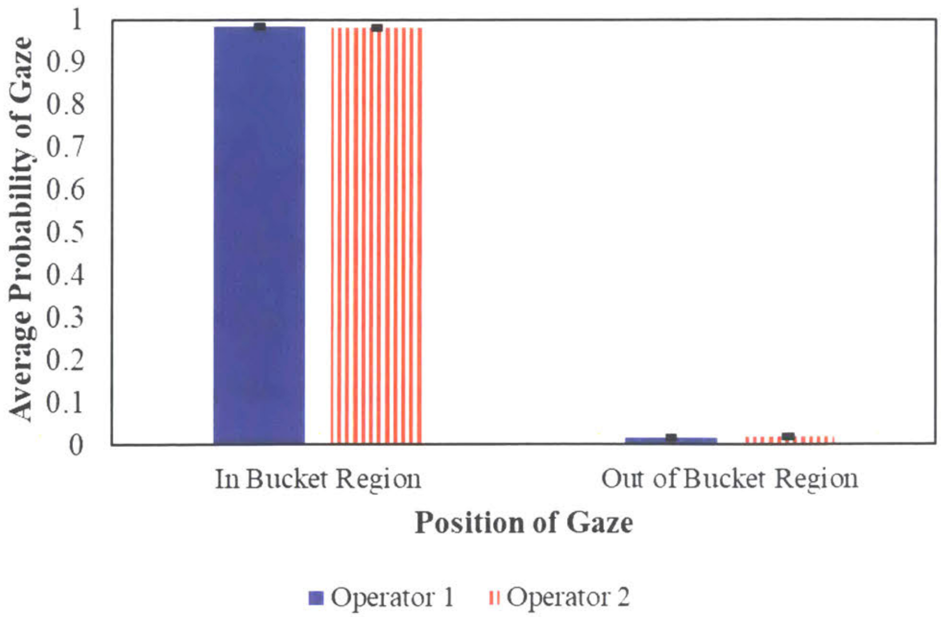

Figure 5-1: The distribution of gaze in and out of the bucket region shows that a key region in a typical operator field of view is the area around the bucket. Operators 1 and 2 fixated on the bucket an average of 98.4 ± 0.5 % (95% C.I.) and 98.0 ± 0.5 % (95% C.I.) respectively.

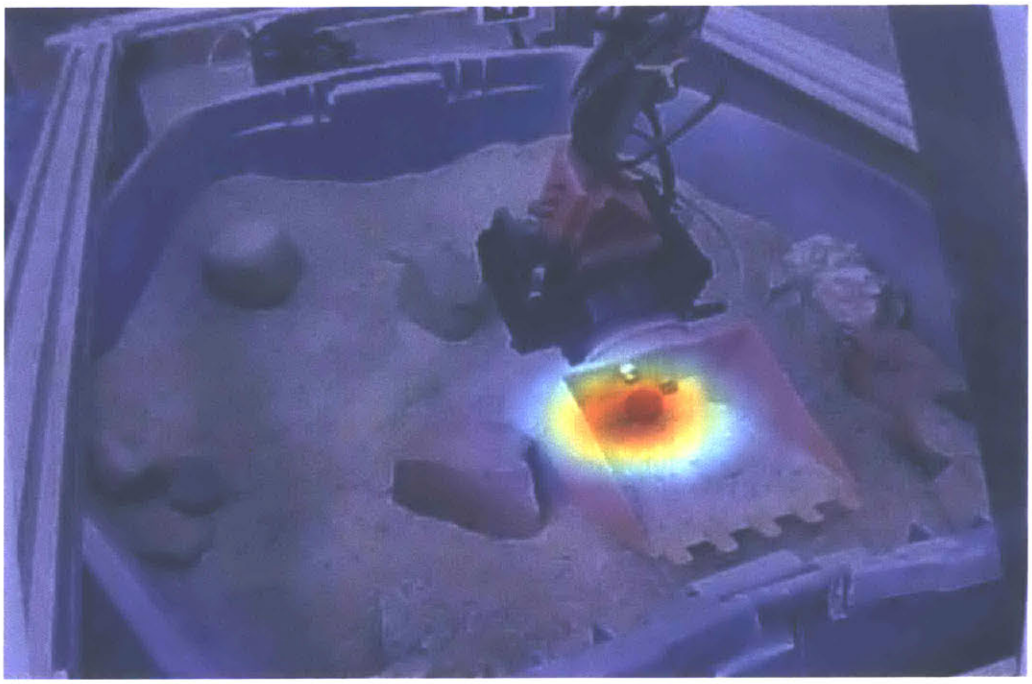

Figure 5-2: A snapshot from the first gaze tracking experiment where the operator was asked to manipulate the bucket in a digging motion given a typical field of view of an excavation operator. The outline of the soil terrain is tracked as a "surface" using markers, while a heat map is superimposed on the surface after image processing. The heat map outlines the most concentrated areas of the operator's gaze in red. In this field of view, the operators had consistent interest in the bucket region.

5.2 Gaze Tracking – Identifying Key Regions within the Bucket

In the second gaze tracking experiment set up, the subject's field of view was limited to the bucket and the soil closely surrounding it to further identify specific regions of interest. This second gaze tracking experiment first further corroborates that the key region of interest is in the bucket. With a closer range field of view for this experiment, an average of 85.5 ± 2.1 % of fixations tracked across all trials and operators, with 95% confidence, were centered in the region where the bucket was. Figure 5-3 shows the distribution of gaze in and out of the bucket region.

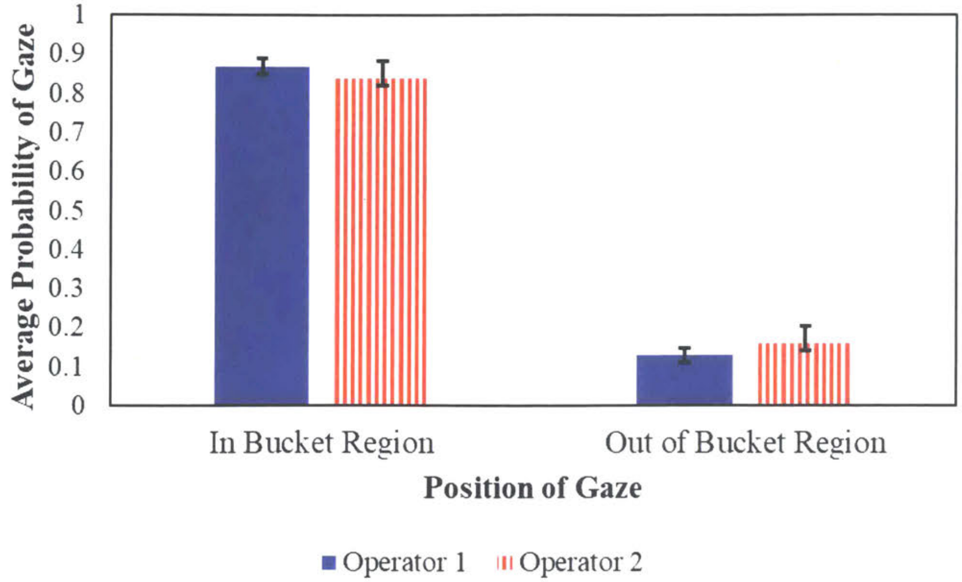

Figure 5-3: The distribution of gaze in and out of the bucket region shows that a key region in a typical operator field of view is the area in the bucket. Operators 1 and 2 fixated on the bucket an average of 87 ± 3 % (95% C.I.) and 84 ± 2 % (95% C.I.) respectively.

Using Pupil's surface tracker as described in section 3, the areas within the bucket with the most gaze concentration throughout the digging task were preliminarily identified. The operators' consistent interest in the bucket region is visually depicted in Figure 5-4 where the gaze distribution graph is red over the bucket.

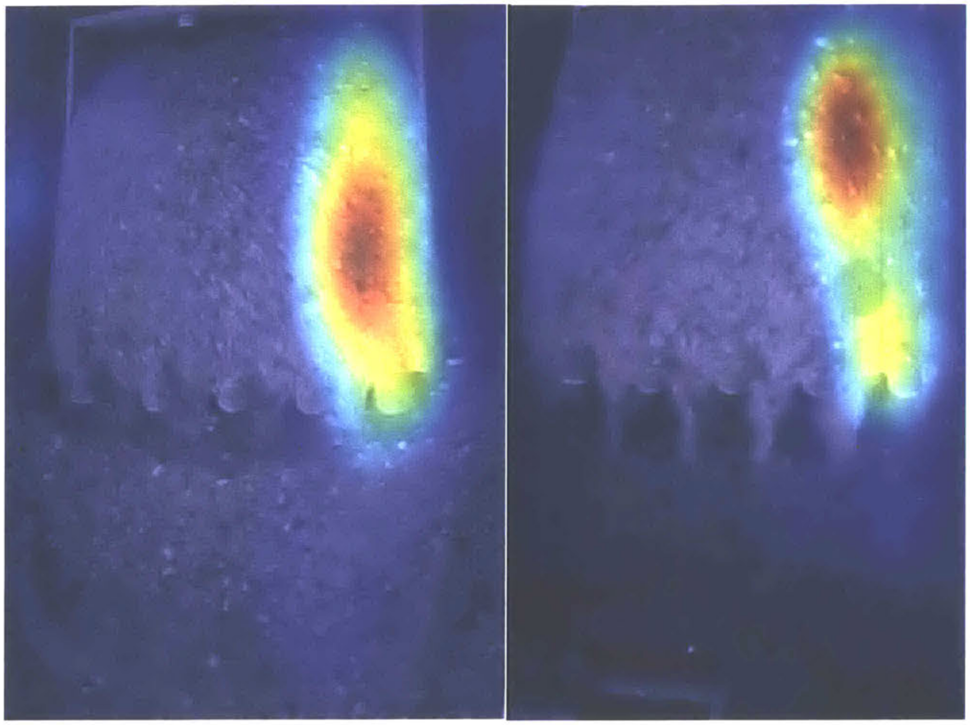

Figure 5-4: Snapshots from the second gaze tracking experiment where the operator was asked to manipulate the bucket given only a view of the bucket. The left and right images show a trial from operators 1 and 2 respectively. The heat map highlights the most concentrated areas of the operator's gaze, which is concentrated on the right-hand side of the bucket for most trials.

To further investigate whether certain regions of the bucket were more of interest to the operators than others, the bucket was divided into quadrants. Figure 5-5 shows the distribution of gaze in the four different quadrants. Quadrant four, the right-hand side of the bucket closest to the opening had the highest concentration of fixations with an average of 68.6 ± 8.2 % of fixations tracked across all trials and operators, with 95% confidence. The concentration of gaze on the right-hand side of the bucket may be due to habits in the operator, but the results are inconclusive. Further tests with more operators should be done to investigate the causes.

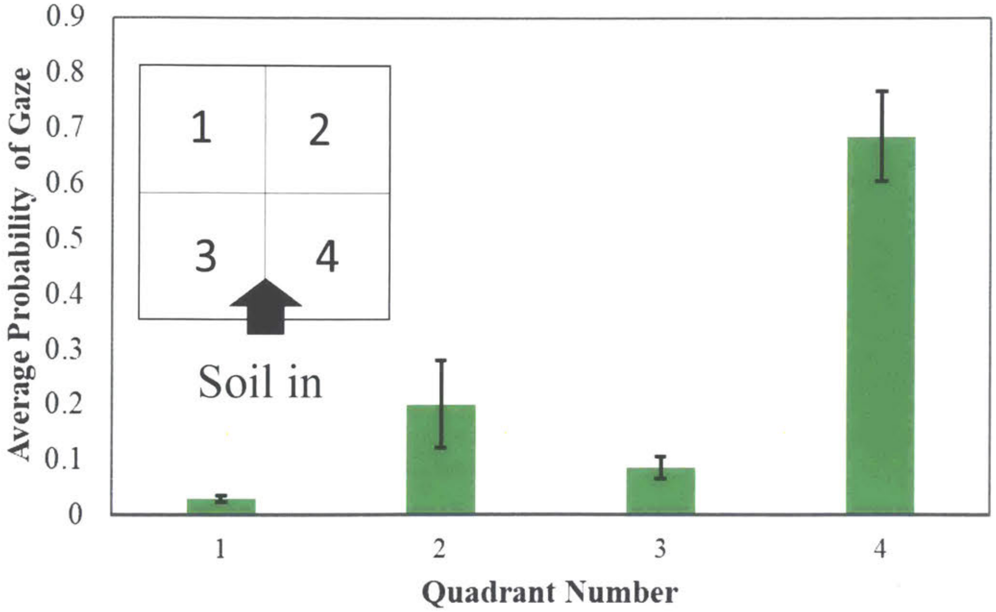

Figure 5-5: The distribution of gaze in four quadrants of the bucket shows that the operator's gaze is typically concentrated in the right-hand quadrant closest to the opening of the bucket.

5.3 Gaze Tracking – Identifying Key Features

The fixation locations were investigated in detail, and an average of 73.5 ± 8.3 % of all fixations were centered around the soil-bucket interface. Figure 5-6 shows a qualitative example of a fixation centered around the soil-bucket interface.

Figure 5-6: A snapshot of a fixation that is centered around the interface between the bucket and the soil.

5.4 Identifying Key Features using Convolutional Neural Networks (CNN)

Having created a dataset as described in section 4, U-Net was trained to recognize the soil-bucket interface in a digging image. Figure 5-7 shows two samples from the train and test sets before and after training. The model identified the interface between the soil and the bucket for all images in the test set.

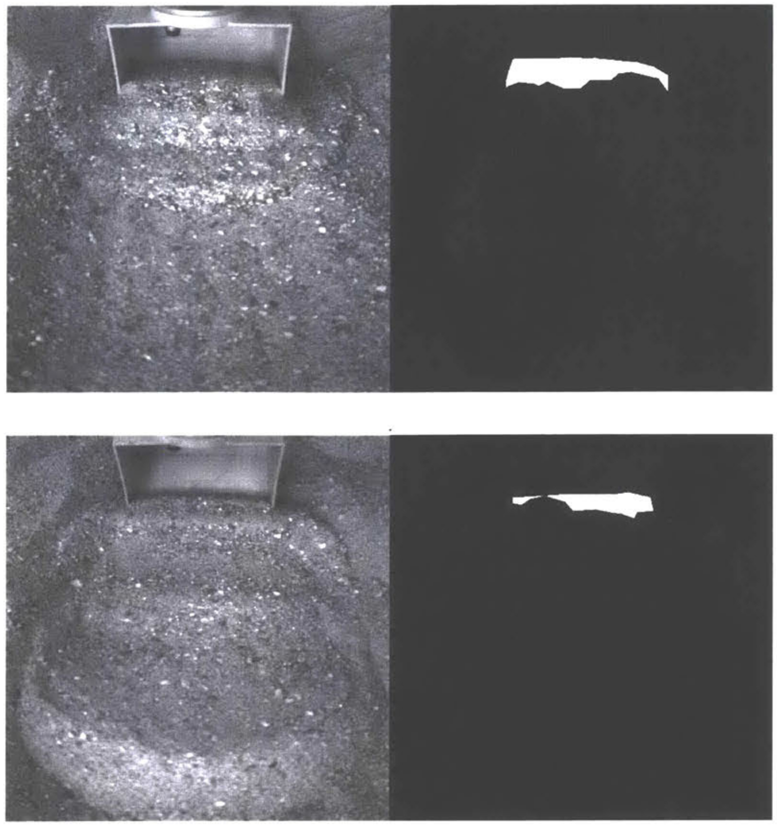

Figure 5-7: Identified soil-bucket interfaces before and after training. The top two images are training labeled dataset examples. The bottom image shows a test sample after the model was trained.

After training the model and testing its capability, the model was used to identify soil-bucket interfaces in gaze tracking experiment videos from section 5.3. The beginning phase of a dig, right before the "picking up" motion was extracted from gaze tracking videos along with gaze positions that correspond to the video segment. After cropping the video where only the bucket was visible, each frame was passed through the trained model to identify the bucket-soil boundary. Figure 5-8 shows an example of a segmentation using a frame that the model has never seen before, indicating its high accuracy. Given a known gaze position for that frame, the position was verified to see whether or not it was in the region of interest, the bucket-soil interface.

Figure 5-8: The segmentation from the U-Net model overlaid on a frame from a gaze tracking video. The segmentation is colored orange for clarity on segmented areas. The model has never seen this frame before, and it is a relatively low quality frame compared to the original images that the model was trained on.

Obtaining the probability that the gaze position was in the segmented bucket-soil interface, the percentage of gaze centered around the soil-bucket interface across two operators for 20 trials was an average of 72.1 ± 5.2 % (95% C.I.), indicating a strong correlation between the soil-bucket interface and human operator attention during the early stage of digging. The average percentage of gaze also is only 1% lower than the average probability of fixations cited in section 5.3.

6. Conclusion

Motivated by the need for automation in excavation, this paper described a methodology in which key regions and features of interest to a human operator can be identified to inform control decisions in the future. The results highlight the key regions, inside the bucket and at the opening of the bucket, and a key feature, the interface between the soil and the bucket, that human operators tend to focus on during a digging task.

In future iterations towards full automation, this methodology of employing gaze tracking to identify features can be applied to an input space of a controls algorithm. For example, in digging, it is useful to understand how operators visually decide the transition point between digging and picking up. Because the system is in constant motion and densities can differ from soil to soil, it is challenging to detect the soil collected in the bucket using force calculations. It is thus very useful to conduct visual analyses, such as optical flow [21], to measure the volumetric flow of soil into the bucket. By comparing the computation time of a controls algorithm of a digging to picking up task while incorporating optical flow with and without the key regions and features identified in this paper, the effects of our method can be measured. The expected outcome is that the computation time is decreased with similar or improved accuracy as the region to conduct real-time visual analysis is isolated.

Furthermore, this methodology can be expanded to apply to many other excavation tasks, such as truck-loading and flattening, as well as other operations in other industries with similar conditions.

7. References

[1] Kawamura, K., Kamada, S., and Tochizawa, M., 1994, "Development of Intelligent Hydraulic Excavator - HYPER GX Series," SAE International.

[2] Saeedi, P., Lawrence, P. D., Lowe, D. G., Jacobsen, P., Kusalovic, D., Ardron, K., and Sorensen, P. H., 2005, "An Autonomous Excavator with Vision-Based Track-Slippage Control," IEEE Trans. Control Syst. Technol., **13**(1), pp. 67–84.

[3] Zolynski, G., Schmidt, D., and Berns, K., 2014, "Safety for an Autonomous Bucket Excavator During Typical Landscaping Tasks," *New Trends in Medical and Service Robots: Challenges and Solutions*, A. Rodić, D. Pisla, and H. Bleuler, eds., Springer International Publishing, Cham, pp. 357–368.

[4] D. Nease, A., and F. Alexander, E., 1993, "Air Force Construction Automation/Robotics," Autom. Constr., **3**, pp. 121–130.

[5] Burks, B. L., Killough, S. M., and Thompson, D. H. (Oak R. N. L., 1992, "Remote Excavation Using the Telerobotic Small Emplacement Excavator," Trans. Am. Nucl. Soc. U. S., **66**.

[6] Huang, X., and Bernold, L. E., 1994, "Control Model for Robotic Backhoe Excavation and Obstacle Handling," ASCE, pp. 123–130.

[7] Lever, P. J. A., Wang, F.-Y., and Chen, D., 1994, "Intelligent Excavator Control for a Lunar Mining System," ASCE, pp. 97–105.

[8] Rocke, D. J., 1996, "Control System for Automatically Controlling a Work Implement of an Earthworking Machine to Capture Material."

[9] Salcudean, S. E., Tafazoli, S., Lawrence, P. D., and Chau, I., 1997, "Impedance Control of a Teleoperated Mini Excavator," *1997 8th International Conference on Advanced Robotics. Proceedings. ICAR '97*, pp. 19–25.

[10] Krause, A., "Synthesis of Tactical Plans For Robotic Excavation," Robot. Inst. Carnegie Mellon Univ.

[11] Cannon, H., and Singh, S., 1999, "Models for Automated Earthmoving," Springer-Verlag London, Sydney, NSW, Australia, pp. 163–172.

[12] Stentz, A., Bares, J., Singh, S., and Rowe, P., 1999, "A Robotic Excavator for Autonomous Truck Loading," Auton. Robots, **7**(2), pp. 175–186.

[13] Rowe, P., and Stentz, A., 1997, "Parameterized Scripts for Motion Planning," *Proceedings of the 1997 IEEE/RSJ International Conference on Intelligent Robot and Systems. Innovative Robotics for Real-World Applications. IROS '97*, pp. 1119–1124 vol.2.

[14] Rowe, P. S., 1999, "Adaptive Motion Planning for Autonomous Mass Excavation," Ph.D., Carnegie Mellon University.

[15] Gang, G., Bo, L., Xuehui, W., Da, Y., and Haibo, Q., 2012, "Study of the Operator Model for Robotic Excavation," *2012 IEEE Symposium on Electrical Electronics Engineering (EEESYM)*, pp. 425–429.

[16] Schwehr, J., and Willert, V., 2017, "Driver's Gaze Prediction in Dynamic Automotive Scenes," *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–8.

[17] Barz, M., and Sonntag, D., 2016, "Gaze-Guided Object Classification Using Deep Neural Networks for Attention-Based Computing," *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16*, ACM Press, Heidelberg, Germany, pp. 253–256.

[18] Barz, M., Daiber, F., and Bulling, A., 2016, "Prediction of Gaze Estimation Error for Error-Aware Gaze-Based Interfaces," *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA '16*, ACM Press, Charleston, South Carolina, pp. 275–278.

[19] Ronneberger, O., Fischer, P., and Brox, T., 2015, "U-Net: Convolutional Networks for Biomedical Image Segmentation," ArXiv150504597 Cs.

[20] Acuna, D., Ling, H., Kar, A., and Fidler, S., 2018, "Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++," ArXiv180309693 Cs.

[21] Krause, A., "An Iterative Image Registration Technique with an Application to Stereo Vision (DARPA)," Robot. Inst. Carnegie Mellon Univ.