Point Cloud Capture and Segmentation of Animal Images using Classification and Clustering

Justin Petluk
justin.petluk@uleth.ca
Department of Mathematics and Computer Science
University of Lethbridge
Lethbridge, Alberta, Canada

Wendy Osborn
wendy.osborn@uleth.ca
Department of Mathematics and Computer Science
University of Lethbridge
Lethbridge, Alberta, Canada

ABSTRACT

Measuring characteristics of animals in the wild is not always possible, due to their demeanour and lack of human contact. Remote capture and processing methods, including the segmentation of animal data into relevant body parts, are required. Existing solutions are either costly or too cumbersome to use in the wild. This study explores the use of RGB depth (RGB-D) cameras for data capture of a target animal from a distance. In addition, this study explores the extraction and segmentation of the resulting animal data into point clouds, and the creation of machine learning models for the automated segmentation of this data. Results of this study, including an experimental evaluation, demonstrate the feasibility of utilizing RGB-D cameras for animal data capture, and that classification outperformed clustering for automated animal data segmentation.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Clustering and classification**; *Data cleaning*; • **Applied computing** → **Psychology**.

KEYWORDS

point cloud, segmentation, classification, clustering

ACM Reference Format:
Justin Petluk and Wendy Osborn. 2021. Point Cloud Capture and Segmentation of Animal Images using Classification and Clustering. In *1st ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility (HANIMOB'21), November 2, 2021, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3486637.3489485

1 INTRODUCTION

Measuring the physical characteristics of animals in the wild has several applications, such as the study of growth patterns. However, in many cases it is difficult to measure the animals directly, either due to their size and demeanour, or since they are not habituated to human contact. Therefore, there is value in the development of methods that enable accurate, remote measurement of physical characteristics such as size [7] or mass [2]. One solution is to use lasers to project light onto a target animal from a known separation distance before photographing, which allows the projected light source to serve as a calibration point for subsequent measurements. Although shown to be useful [1], the object of interest must be parallel to the focal plane of the camera, and the light source must be recalibrated continuously, leading to a cumbersome setup. Another solution is LIDAR (Light Detection and Ranging) technology, a remote sensing method for collecting and reconstructing spatial data remotely. LIDAR also has been shown to be useful in controlled environments, including estimating human joint locations [12], the length and height of cattle [7], and the weight of primates [2]. Although LIDAR can be used in uncontrolled environments with promising results, it is costly and has limited portability. Another option is an RGB-D (RGB depth) camera, such as Intel® RealSense™ D415 [1]. Both affordable and compact, it captures depth information on a target from up to 10 metres away in a short time period, both indoors and outdoors [3], as long as exposure is taken into consideration [5]. The addition of depth measurements helps reduce the errors that are introduced from remote capturing, which addresses the issue of requiring that the target is parallel to the camera. Multiple cameras can be used to capture data simultaneously from different perspectives, which improves accuracy [4] without interference from each other [5].

This study explores the initial steps of reliable, practical and cost-effective data collection and extraction in field environments, for supporting non-invasive techniques. Specifically, this study explores the following questions: 1) How well can multiple RGB-D cameras be used for effectively capturing a target animal in a controlled outdoor environment, 2) How should the RGB-D camera data be processed into point cloud representations of the target animal, as well as the body parts that make up the animal? and 3) Which classification and clustering models are ideal for body part identification and segmentation of the target animal?

2 STRATEGY

This section presents our strategy for the segmentation of a point cloud using classification and clustering. First, preliminary information on point clouds, classification and clustering are provided that is relevant to this study. Then, we present our strategy, including: data capture, data preparation, neural network model creation, segmentation using classification, and segmentation using clustering. A point cloud is a set of points in three-dimensional space that collectively has an arbitrary shape. Along with additional information such as colour and surface normals, a point cloud represents a complex object (e.g. animal).

[1] https://www.intelrealsense.com/depth-camera-d415/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HANIMOB'21, November 2, 2021, Beijing, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-9122-1/21/11…$15.00
https://doi.org/10.1145/3486637.3489485

Classification [6] is a two-step process. First, a model is created using class-labelled data. Then, this model is used to predict the class label for unlabelled data. For example, a model can be trained using point clouds of an animal and its relevant parts. Then, this model can be used to predict the relevant parts of other instances of the same animal. We use the PointNet++ [9] and PointCNN [8] neural network architectures for classification model creation and prediction. PointNet++ creates and learns various degrees of example point cloud densities in order to adapt to a sparse set of data. This data is then passed on and the sampled data is interpolated and combined in order to predict per-point segmentation scores. PointCNN also learns various degrees of point clouds but combines those features with the original features and then uses convolutional layers. To predict per-point segmentation labels, PointCNN uses deconvolutional layers, receiving data from each respective convolution to predict per-point segmentation scores.

Clustering [6] uses mathematical or statistical measures to separate the data into meaningful groups (e.g. relevant parts of an animal). Although data is not required to be labelled in advance, other pre-defined values (e.g. number of clusters) must be provided in many cases. We use the Constrained Planar Cuts (CPC) [11] and Region Growing [13] strategies for clustering our point cloud data. Constrained Planar Cuts clusters points based on colour, normals and distance. These cluster centroids are then connected, and the faces define a convex or concave contour. A plane is fit to cut along concave groups both locally and globally. Region Growing Segmentation groups randomly selected points based on smoothness, curvature and colour. Thresholds are used to control which regions should no longer merge.

2.1 Data Capture

To create several point clouds of our target animal, we used an *Iterative Capture* strategy to capture several images from various distances, and both the left- and right-hand sides of the animal. We used two Intel® RealSense™ D415 RGB-D cameras, a taxidermy raccoon as our animal, and an outdoor setting. The Iterative Capture strategy is carried out as follows. First, the cameras are placed 91.5cm (3 feet) apart, and six metres away from the left-hand side of the animal. After several initial images are captured from both cameras, they are moved 30 centimetres closer to the animal. After several more images are captured, the cameras are moved another 30 centimetres closer to the animal. This process is repeated until a distance of 50 centimetres from the animal is reached. Then, the same Iterative Capture process is repeated from the right-hand side of the animal. In total, 236 images were acquired. From these images, the best 43 images were chosen from each of the left- and right-hand sides, for a total of 86 images. The selected images provided the best perspectives of the animal, with distinguishable body parts that will best serve the following stages of this work. The remaining 150 images were discarded due to issues with their quality. Some of these issues include: 1) indistinguishable body parts (e.g. one arm could not be distinguished from the body), 2) low image quality due to too few point samples acquired, and 3) significant amounts of noise introduced into the image. In particular, images captured between the 3 and 6 metres range proved to be the least reliable. We feel that the 10-metre capture range stated for the camera also depends on the size of the animal, as a smaller animal may be harder to capture clearly from a farther distance. The Librealsense software (provided with the camera) provides post-processing support for spatial, temporal, and resolution reduction of images. The default spatial filter improved the edge preservation of objects within an image, while the default temporal filter improved the smoothness of the image by removing noise. Finally, a decimation filter provided a 4-to-1 resolution reduction, where images were reduced from 1280x720 to 640x360. These post-processing steps both improve the quality of the images, as well as reduce their size, for the remaining stages of this work. In particular, having higher-resolution images may result in significantly increased execution times for certain tasks, such as model training for classification.

2.2 Data Preparation

Once a candidate set of images is selected, they are transformed into a collection of point clouds. Figure 1 presents the process of data preparation for one image (i.e. point cloud scene). The steps *Segment Racoon from Scene* and *Segment Body Parts* are performed first. They generate all required inputs for *Network Training*, *Classification-based Segmentation* and *Clustering-based Segmentation*.

For the first step, a point cloud of the target animal is extracted from its containment image and saved into a file. For the second step, the target animal is segmented further into its components - the head (including the neck and ears), the body (including the thighs and legs), the tail, and the arms (the left arm including the shoulder, and the right arm up to the elbow). Each component is also saved into a separate point cloud file. CloudCompare [2] is used for both of these segmentation tasks. CloudCompare was chosen for its ease of setup and use. Segmentation of both the image and the target animal itself simply requires the drawing of a polygon around the points of interest, followed by applying a polygon cut tool to extract the selected points. One limitation of CloudCompare is the lack of a tool for assigning labels to individual points or sets of points. Therefore, manual labelling of the points that make up individual parts of the target animal is required.

In this work, we adopt the following semi-automated approach for point labelling. First, an identifying label - head, body, arm, or tail - is assigned manually to each set of points that represent a body part. Then, using this information, and an *identifying label* $\rightarrow$ *class label* mapping, the class label for each individual point of the body part is assigned automatically via a script. To assign an identifying label to a point cloud that represents a body part, we use the filename of the point cloud for this purpose (e.g., Body.ply is the filename for a point cloud that represents the body of the animal). Then, using the information in the filename, and the mapping (head, body, arm, tail) $\rightarrow$ (1,2,3,4), each point is assigned a class label.

Finally, the points on the original point cloud of the target animal are class labelled. This could be done by iterating through the points, searching the corresponding limbs for the matching point, and assigning the same class label. However, computationally this process is costly. Therefore, the approach we took was to re-construct the animal point cloud using its class-labelled body parts. First, during the process of class labeling, all body parts that belong to a specific animal are placed together in a folder, with the name of the folder

[2] http://www.cloudcompare.org/

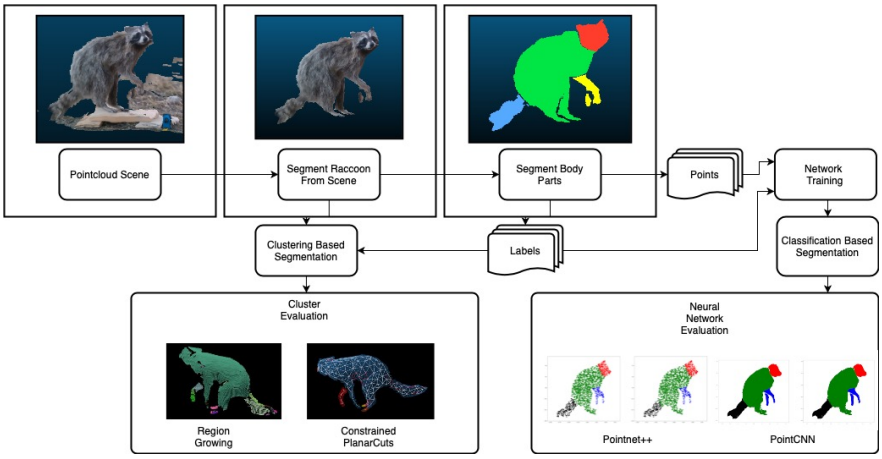

Figure 1: Flowchart of Strategy

corresponding to the filename of the original point cloud for this animal. The approach fetches the point cloud for each body part in the order of their class label, and concatenates them into one list. At the same time, a second list of IDs is created, which is a list of class labels where each class label corresponds to the point in the same location in the Point list. For example, assume we have the following point clouds for an animal: Head [p1,p2], Body [p3,p4,p5,p6] and Arm [p7,p8]. The ID and Point lists for the reconstructed animal are: ID [1,1,2,2,2,2,3,3] and [p1,p2,p3,p4,p5,p6,p7,p8].

2.3 Model Training and and Segmentation

Once the captured data is processed, the classification and clustering models must be configured and/or trained, in order to be able to perform automatic segmentation of animal point cloud data. For PointCNN and PointNet++, the reconstructed versions of the animal point clouds are required as input for training. For PointNet++, the point normals were also used, which were calculated using the Point Cloud Library's surface normal estimation method [10]. Only 2/3rds of the collection of reconstructed animal point clouds were used for training PointCNN and PointNet++. The remainder of the collection was used for evaluation purposes, which is discussed in the next section. Once the models are trained, any animal point clouds can be segmented into its parts automatically with the model. For CPC and Region Growing – the original versions of the animal point clouds (before segmentation into parts) are required as input. Note that CPC also uses the colours of the animal in the clustering process, while Region Growing does not. For both models, the output is the animal segmented into its parts.

3 EVALUATION

In this section, we present the methodology, results and discussion of our experiments. For classification, PointNet++ and PointCNN were evaluated using 1/3 of the dataset of testing and 2/3 of the dataset for training. For testing accuracy, the Mean Jaccard Index similarity measure was used at the end of training, which is the average of the Jaccard Index values for predicting the head, body, arms and tail. For training accuracy, the percentage of correctly labelled points was calculated throughout training. It must be noted that PointNet++ evaluates over epochs, while PointCNN evaluates over iterations with 12 iterations equal to an epoch and 8 point clouds per iteration. For clustering, CPC and Region Growing were evaluated by using the labelled animal segments as ground truth, for comparison against what was predicted from the original animal point cloud. For all predicted segmentations, homogeneity (i.e. how well points exist within the class they belong to), completeness (i.e., how well every cluster point exists in a class), and V-measure (i.e. ratio of homogeneity/completeness) were measured.

3.1 Results and Discussion

Overall, both PointNet++ and PointCNN performed quite well on the given set of animal point clouds. For testing accuracy, the mean Jaccard Index for PointNet++ was 0.95 and for PointCNN was 0.94. The training accuracy results are depicted in Figures 2 and 3. At the end of training both PointNet++ and PointCNN achieved training accuracies close to 1. The difference is in how quickly they converged on that accuracy value. PointCNN achieved a high learning accuracy almost immediately while PointNet++ took longer to reach the same accuracy. In addition, PointCNN was better at predicting the edges of each body part, while PointNet++ struggled with edge prediction. Both Region Growing and CPC performed poorly when compared to the classification results. Figures 4 and 5 depict these results. With a more consistent V-measure, Region Growing outperforms CPC. However, CPC achieved better completeness. With respect to classification, although in general the varying body parts were identified, the largest source of error was found to be in the ability to accurately label along the edges of each body part. In particular, labelling along the arms or tail where they met with the body led to the most extreme sources of error. In these cases the size of the body was over-estimated. These errors could be the result of user labeling errors, or other issues in the training or testing data. One additional challenge was in how testing accuracy was calculated. Although obtaining a final testing accuracy was possible, PointNet++ also calculated testing accuracy throughout the training process, but PointCNN did not. It would have been

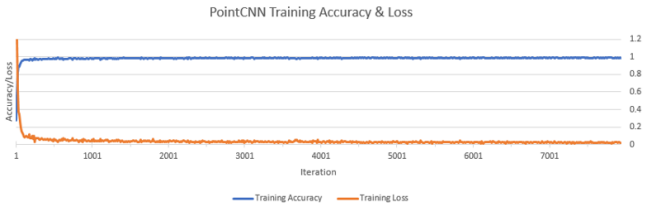

Figure 2: PointCNN

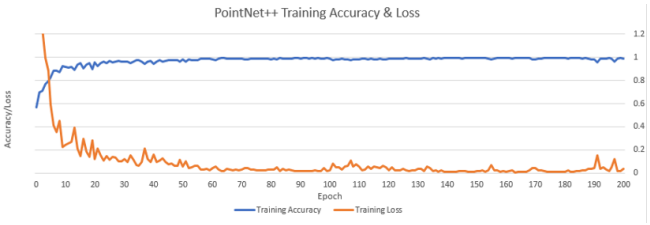

Figure 3: Pointnet++

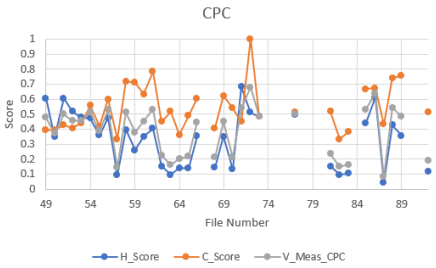

Figure 4: CPC

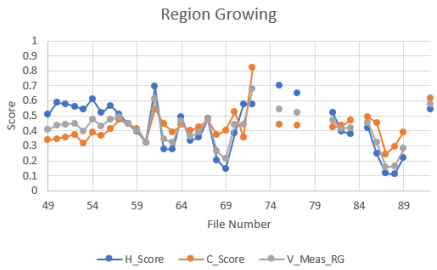

Figure 5: Region Growing

ideal to carry out this comparison at the epoch level. Although clustering did not perform as well as classification, some success was found by identifying the individual results. As seen in both curves, results from the initial few point cloud animals were significantly better than the results that came later on. This is likely a reflection of the variation of the point clouds over the entire dataset, while the clustering parameters were set once and maintained throughout the entire clustering process. This reflects the nature of a large group of clustering methods, which still continues to be a hurdle. In addition to having more accurate datasets, clustering strategies would need to be able to self-tune to adapt to changes in the training data.

4 CONCLUSION AND FUTURE WORK

This study explores reliable, practical and cost-effective data collection and extraction in field environments. The answers to our research questions are: 1) Multiple RGB-D cameras show promise in capturing an animal in a controlled outdoor environment, with the best results coming from within 3 metres, 2) We use a semi-automated approach to segment the animal into its parts, followed by a re-construction process to form the entire animal point cloud, 3) We found that both classification models worked well in segmenting animals into their parts (although challenges existed at the edges), while clustering did not perform as well overall. Future work directions include evaluating our strategy using other types of animals and species, in the wild, and improving our strategy by performing data extraction after segmentation.

ACKNOWLEDGMENTS

Thanks to Louise Barrett and Peter Henzi from the Barrett/Henzi Lab, University of Lethbridge for providing the resources and support to make this project possible!

REFERENCES

[1] Andreas Berghänel, Oliver Schülke, and Julia Ostner. 2015. Locomotor play drives motor skill acquisition at the expense of growth: a life history trade-off. *Science Advances* 1, 7 (2015), 8 pages.
[2] Charlotte A. Brassey and William I. Sellers. 2014. Scaling of convex hull volume to body mass in modern primates, non-primate mammals and birds. *PLoS One* 9, 3 (2014), 12 pages.
[3] Monica Carfagni, Rocco Furferi, Lapo Governi, Chiara Santarelli, Michaela Servi, Francesca Uccheddu, and Yary Volpe. 2019. Metrological and Critical Characterization of the Intel® D415 Stereo Depth Camera. *Sensors* 19, 3 (2019), 20 pages.
[4] Stefan K. Gehrig and Uwe Franke. 2007. Improving stereo sub-pixel accuracy for long-range stereo. In *Proceedings of the 11th IEEE International Conference on Computer Vision*. 1–7.
[5] Anders Grunnet-Jepson, John N. Sweetser, and John Woodfill. [n. d.]. Best-known methods for tuning Intel® RealSense™ D400 Depth Cameras for Best Performance. https://www.intel.com/content/dam/support/us/en/documents/emerging-technologies/intel-realsense-technology/BKMs_Tuning_RealSense_D4xx_Cam.pdf (last accessed May 2021).
[6] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data mining concepts and techniques (3rd ed.)*. Morgan Kaufmann Publishers, Waltham, Mass.
[7] Lvwen Huang, Shuqin Li, Anqi Zhu, Xinyun Fan, Chenyang Zhang, and Hongyan Wang. 2018. Non-contact body measurement for Qinchuan cattle with LIDAR sensor. *Sensors* 18, 9 (2018), 21 pages.
[8] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. 2018. PointCNN: Convolution on X-transformed Points. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*. 11 pages.
[9] Charles R. Qi, Li Yi, Hao. Su, and Leonidas J. Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Proceedings of the 31st Conference on Neural Information Processing Systems*. 5105–5114.
[10] Radu B. Rusu and Steve Cousins. 2011. 3D is here: Point Cloud Library. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation*. 4 pages.
[11] Markus Schoeler, Jeremie Papon, and Florentin Wörgötter. 2015. Constrained Planar Cuts - Object Partitioning for Point Clouds. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*. 5207 – 5215.
[12] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*. 1297–1304.
[13] Anh-Vu Vo, Linh Truong-Hong, Debra F. Laefer, and Michela Bertolotto. 2015. Octree-based region growing for point cloud segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing* 104 (2015), 88–100.