



7-13-2017

Examining the Use of Amazon's Mechanical Turk for Edge Extraction of the Occlusal Surface of Fossilized Bovid Teeth

George K. Thiruvathukal
Loyola University Chicago, gkt@cs.luc.edu

Gregory J. Matthews
Loyola University Chicago, gmatthews1@luc.edu

Maxwell P. Luetkemeier
Loyola University Chicago

Juliet K. Brophy
Louisiana State University at Baton Rouge

Follow this and additional works at: https://ecommons.luc.edu/cs_facpubs

 Part of the [Computer Sciences Commons](#), and the [Paleobiology Commons](#)

Recommended Citation

Thiruvathukal, George K.; Matthews, Gregory J.; Luetkemeier, Maxwell P.; and Brophy, Juliet K.. Examining the Use of Amazon's Mechanical Turk for Edge Extraction of the Occlusal Surface of Fossilized Bovid Teeth. *PLoS One*, 12, 7: 1-18, 2017. Retrieved from Loyola eCommons, Computer Science: Faculty Publications and Other Works, <http://dx.doi.org/10.1371/journal.pone.0179757>

This Article is brought to you for free and open access by the Faculty Publications and Other Works by Department at Loyola eCommons. It has been accepted for inclusion in Computer Science: Faculty Publications and Other Works by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.



This work is licensed under a [Creative Commons Attribution 4.0 International License](#).
© Gregory J. Matthews, et al. 2016

RESEARCH ARTICLE

Examining the use of Amazon's Mechanical Turk for edge extraction of the occlusal surface of fossilized bovid teeth

Gregory J. Matthews^{1*}, George K. Thiruvathukal², Maxwell P. Luetkemeier¹, Juliet K. Brophy³

1 Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL, United States of America, **2** Department of Computer Science, Loyola University Chicago, Chicago, IL, United States of America, **3** Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA, United States of America

* gmatthews1@luc.edu



Abstract

In order to reconstruct environments associated with Plio-Pleistocene hominins in southern Africa, researchers frequently rely upon the animals associated with the hominins, in particular, animals in the Family Bovidae. Bovids in southern Africa are typically identified by their teeth. However, identifying the taxon of a bovid tooth is challenging due to various biasing factors. Furthermore, inaccurate identification of fossil bovids can have significant consequences on the reconstructed paleoenvironment. Recent research on the classification of bovid fossil teeth has relied on using elliptical Fourier analysis to summarize the shape of the outline of the occlusal surface of the tooth and the resulting harmonic amplitudes. Currently, an expert in the field must manually place landmarks around the edges of each tooth which is slow and time consuming. This study tests whether it is possible to crowdsource this task, while maintaining the necessary level of quality needed to perform a statistical analysis on each tooth. Amazon Mechanical Turk workers place landmarks on the edge of the tooth which is compared to the performance of an expert in the field. The results suggest that crowdsourcing the digitization process is reliable and replicable. With the technical aspects of digitization managed, researchers can concentrate on analyzing and interpreting the data.

OPEN ACCESS

Citation: Matthews GJ, Thiruvathukal GK, Luetkemeier MP, Brophy JK (2017) Examining the use of Amazon's Mechanical Turk for edge extraction of the occlusal surface of fossilized bovid teeth. PLoS ONE 12(7): e0179757. <https://doi.org/10.1371/journal.pone.0179757>

Editor: Sergi Lozano, Institut Català de Paleoecologia Humana i Evolució Social (IPHES), SPAIN

Received: December 5, 2016

Accepted: June 2, 2017

Published: July 13, 2017

Copyright: © 2017 Matthews et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data are available in the GitHub repository (<https://github.com/gjm112/MTurkTeeth>).

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Reconstructing past environments associated with early hominins is essential for understanding human evolution and is valuable for identifying habitat preferences, diet, and ecological relationships between hominins and other species. In order to reconstruct past environments, paleoanthropologists commonly rely on the animals that are found associated with the hominins. Animals in the Family Bovidae such as antelopes and buffalo are particularly useful for this task due to their strict ecological tendencies [1–3]. In addition, bovids are one of the most common fossils found in southern Africa, in particular isolated teeth. However, identifying

bovid teeth in the fossil record is complicated by biasing factors such as attrition and sex [4]. Overlap exists in the form (i.e. size and shape) of bovid teeth making it difficult to identify the taxon and, therefore, difficult to reconstruct the past environment [4]. The purpose of this study is to demonstrate a reliable, replicable, uncomplicated method for extracting the form of the occlusal surface of bovid teeth which can then be used to identify teeth in the fossil record. Several recent studies have demonstrated that morphometrics is particularly useful for documenting biological shape [5–11]. This new methodology extracts edges by relying on crowdsourcing. The outlines are then used in supervised machine learning techniques in conjunction with elliptical fourier analysis (EFA) [12].

It should be noted that ideally edge extraction of the occlusal surface of these teeth could be performed using automated procedures based on techniques such as those described in [13] or [14]. However, in this specific setting automated methods are difficult to use as these techniques tend to often identify the bottom of a tooth as the edge rather than the actual occlusal surface.

Previously, [1] performed a study to standardize the identification of bovid teeth using EFA. While successful in identifying bovid taxa, the process to extract the outlines was tedious and time consuming. In order to extract the outline of a tooth, an image was imported into a digitizer program, MLmetrics [15], where 60 points were manually placed around the tooth according to a template so as to maintain homology. The points were then exported and analyzed in a fourier analysis program [16]. The study generated occlusal outline information for over 7000 extant and fossil teeth. However, the results could not be easily used to identify fossils from new sites due to the time consuming nature of the process of edge extraction. The present study provides results of an exploratory analysis that employs Amazon's Mechanical Turk platform [17] as a method to crowdsource the edge extraction of bovid teeth.

In this study, the digitized outlines of an expert in the field, the co-author Juliet K Brophy (JKB), are compared with up to three outlines extracted by Amazon Mechanical Turk workers. The results of this preliminary study suggest that crowdsourcing the digitizing process is reliable and replicable. Furthermore, this streamlined process allows for more teeth to be processed in a timely manner, saves the time of researchers from performing technical tasks, and frees them up to focus more of their time on aspects of this project that require expertise, such as analyzing and interpreting the data.

2 Related work

Mechanical Turk [17] was introduced by Amazon.com, Inc. in 2005. As such, there is a relatively limited body of scholarly work exploring the uses of the platform. The projects that task quality assessment, the focus of this study, can be divided into two categories: assessing survey response accuracy and annotating digital images.

2.1 Assessing survey response accuracy

Studies in this category focus on investigating how accurate survey responses are from Mechanical Turk Workers. These analyses aim to answer questions such as: How closely do Mechanical Turk surveys reflect surveys distributed using more traditional methods? [18]; How honest are Mechanical Turk workers in their responses? [19, 20]; and Does Mechanical Turk provide researchers with a more diverse response pool than the mainstay of distributing surveys to college students with the promise of extra credit? [21]

[20] uses Mechanical Turk in order to combine the speed and cost-effectiveness of a simulated study with the authenticity of human behavioral studies when analyzing human cooperation. The study claims that prior to Mechanical Turk and the ability to crowdsource data

collection, most evolutionary models were based on simulations or mathematical algorithms due to the lack of survey labs and a consistent subject pool. With its use, however, researchers can request a task to be done and collect results entirely online much in the same way a simulation study is conducted. With that said, [20] mentions that a major concern of using Mechanical Turk is the lack of control researchers have over their subjects. It is possible, for instance, for subjects to incorrectly answer a question due to a lack of understanding. Additionally, subjects are completely free to leave in the middle of the survey. After conducting a number of experiments, both online and in person, [20] found that these limitations had a very small effect on the results.

In a similar study, [21] conducted an experiment comparing the performance of Mechanical Turk workers versus subjects in a controlled laboratory setting in an acceptability judgment task. The main concern addressed in [21] is that additional noise, introduced by using Mechanical Turk, might detract from the power of the experiment. To help control for this, they introduced a rejection criteria. Mechanical Turk workers were required to be native English speakers, which resulted in a 15% rejection rate. [21], like [20], states that another major concern in the use of Mechanical Turk is the inability to establish whether or not the Turker understood the task, possibly resulting in inaccurate data. It concluded, however, that using Mechanical Turk is comparable to laboratory research as long as a mechanism exists to reject certain responses.

Additional information on testing best practices when using Mechanical Turk in survey research can be found in [19], which evaluates how various factors effect the reliability of responses, and [18], which compares the demographics of Mechanical Turk respondents to national demographics.

2.2 Annotating digital images

This category of Mechanical Turk work evaluates the quality of edge extraction research. Two of the primary works related to this topic include [22] and [23].

[22] explored the use of Mechanical Turk in image classification focusing on techniques for automatically “cleaning” the data sets. They demonstrate that by using multiple methods for measuring the accuracy of annotations they can outperform other methods that rely on a single measure. They also demonstrate that image classification can be performed with high levels of accuracy when using Mechanical Turk workers to extract the edge of images. Further, classification accuracy can be improved by over 7%, by cleaning the data using the techniques considered in this study.

[23] evaluates various annotation techniques with the goal of maximizing quality while minimizing cost. This research used landmark-based edge extraction and a gold standard method of grading. Landmark extraction, or annotation, involves having a Turker place a number of points along the border of an image. Once the outline is extracted, it can be tested for quality against an outline annotated by an expert, which is referred to as the “gold standard” grading technique. While it was not used in this particular study, [23] also mentions grading outlines based on their distance from the mean image produced by multiple Mechanical Turk workers, which may be useful as it eliminates the need for expert tracing.

3 Methods

This exploratory study includes a sample of 96 teeth of known species from four different tribes: Alcelaphini, Bovini, Hippotragini, and Neotragini. These teeth were obtained from the Ditsong Museum (TM) (formerly Transvaal Museum) and the National Museum of Bloemfontein (NMB), South Africa. (Permission to use these specimens was received by JKB from

Table 1. List of extant bovid specimens used in this study from the National Museum, Bloemfontein (NMB) and the Ditsong Musuem (TM) (formerly Transvaal Museum).

Genus	Species	Repository	Specimen Number
UM3			
<i>Connochaetes</i>	<i>taurinus</i>	NMB	64, 12066, 12204, 12475
<i>Damaliscus</i>	<i>dorcas</i>	NMB	8752, 9382, 12159, 12175, M144
<i>Oryx</i>	<i>gazella</i>	NMB	250, 9304, 9330, 12094, 12181
<i>Hippotragus</i>	<i>niger</i>	NMB	177, 178, 183, 232, 893
<i>Syncerus</i>	<i>caffer</i>	NMB	9, 12, 16
UM2			
<i>Hippotragus</i>	<i>equinus</i>	NMB	191, 196, 887
<i>Hippotragus</i>	<i>equinus</i>	TM	AZ 1133
LM3			
<i>Alcelaphus</i>	<i>buselaphus</i>	NMB	6022, 8715, 8763, 12199, 12215
<i>Raphicerus</i>	<i>campestris</i>	NMB	8730, 9343, 9438, 9761, 9787
<i>Pelea</i>	<i>capreolus</i>	NMB	9446, 6878
<i>Pelea</i>	<i>capreolus</i>	TM	AZ 479, 10005, 10007
<i>Syncerus</i>	<i>caffer</i>	NMB	1000, 1001, 1002, 8743, 8774
LM2			
<i>Connochaetes</i>	<i>gnou</i>	NMB	12218, 12323, 12399, 12394
<i>Connochaetes</i>	<i>taurinus</i>	NMB	12201, 12204, 12209, 12475, 12476
<i>Alcelaphus</i>	<i>buselaphus</i>	NMB	6022, 12196, 12199, 12215, 12420
<i>Damaliscus</i>	<i>dorcas</i>	NMB	7440, 9384, 12039, 12157
<i>Raphicerus</i>	<i>campestris</i>	NMB	8730, 9438, 9761, 9787, 12169
<i>Pelea</i>	<i>capreolus</i>	NMB	6878, 9446, 9855
<i>Oryx</i>	<i>gazella</i>	NMB	9304, 9335, 12182, 12213, 12352
<i>Hippotragus</i>	<i>niger</i>	NMB	176
<i>Hippotragus</i>	<i>niger</i>	TM	3812, 4251, 13130, 13136
<i>Hippotragus</i>	<i>equinus</i>	NMB	191, 887
<i>Hippotragus</i>	<i>equinus</i>	TM	AZ 2444, AZ 1333, 12072
LM1			
<i>Alcelaphus</i>	<i>buselaphus</i>	NMB	8790, 12195, 12420
<i>Damaliscus</i>	<i>dorcas</i>	NMB	9384, 12039, 12157, 12320, 15155
<i>Hippotragus</i>	<i>niger</i>	TM	3812, 13130, 13138, 13143, 13153

<https://doi.org/10.1371/journal.pone.0179757.t001>

both institutions (i.e. National Museum, Bloemfontein and Ditsong Museum (formerly Transvaal Museum)). Permits are not required to look at extant bovid specimens in South Africa. Therefore, no permits were required for the described study.) The complete repository information is in [Table 1](#). Permission was received from each institute to photograph these specimens. No permits were required for the described study, which complied with all relevant regulations.

We investigated three mandibular molars (LM1, LM2, LM3) and two maxillary molars (UM2, UM3). Details of the data are shown in [Table 2](#). An example of the raw image of a tooth prior to extraction can be seen in the left side of [Fig 1](#). Prior to being digitized by a Turk worker, all of the teeth were scaled to each other.

Table 2. The distribution of tribe by tooth type in the data set.

Tribe	Tooth Type					Total
	LM1	LM2	LM3	UM2	UM3	
Alcelaphini	8	15	5	0	5	33
Bovini	0	0	5	0	5	10
Hippotragini	5	15	0	5	10	35
Neotragini	0	8	10	0	0	18
Total	13	38	20	5	20	96

<https://doi.org/10.1371/journal.pone.0179757.t002>



Fig 1. Raw image of tooth.

<https://doi.org/10.1371/journal.pone.0179757.g001>

3.1 HIT protocol

Amazon states: “A Human Intelligence Task, or HIT, is a question that needs an answer. A HIT represents a single, self-contained task that a Worker can work on, submit an answer, and collect a reward for completing” [24]. Specifically in this setting, the Mechanical Turk worker downloads the image of a bovid tooth in the freeware GIMP (the GNU Image Manipulation Program) [25]. After testing several programs for obtaining the polygon, this program produced the best results. Next, the Mechanical Turk worker selects the *lasso tool* which allows a polygonal selection to be made around the tooth. Once the bounding polygon has been created, the user then cuts and pastes the extracted selection onto a blank canvas. This shape is then filled in with all black using the bucket fill tool in GIMP creating a black and white image of each tooth where the interior of the tooth is black and the background is all white. The resulting file is then saved onto one’s computer and uploaded to the link provided in the HIT.

3.2 Processing the Mechanical Turk output

For every raw image of a tooth considered in this study, Mechanical Turk workers were asked to extract the outline of the occlusal surface in GIMP [25]. This process was repeated 3 times for each tooth. (Mechanical Turk workers were used only to trace images of bovid teeth. No personal information relating to any mechanical Turk worker was collected.) The output from

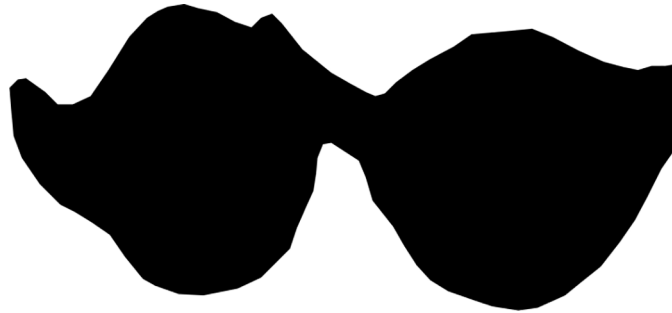


Fig 2. Extracted Occlusal Surface using GIMP performed by expert.

<https://doi.org/10.1371/journal.pone.0179757.g002>

each of the Mechanical Turk workers was then forced to a black and white image using Image-Magick [26]. The expert (JKB) also traced the outline of the occlusal surface of each tooth using the exact methods as the Mechanical Turk workers.

Fig 1 shows an example of a raw image of a tooth, and Fig 2 shows the tracing by an expert using GIMP. The results from Mechanical Turk workers for this specific tooth are shown in Fig 3. Each of these three images corresponds to different tracings of the raw tooth presented in Fig 1. Note that the tracing on the bottom of Fig 3 was not done correctly by the Mechanical Turk worker and needed to be adjusted after the fact to an image that is strictly black and white.

While three separate HITs for each tooth were posted, we were not always able to get three viable tracings. In some cases, no tracing was returned whereas in other cases, tracings were returned but were clearly wrong. This circumstance occurred, for instance, when Turk workers traced around the occlusal surface of teeth that were not the focus of the image, traced some parts of the internal area of the tooth, or returned a .xcf file (a GIMP file) instead of the .jpg that was requested in the HIT.

Once the black and white images are collected and processed, they can be read into R using the “import_jpg” function from the Momocs package [27]. This function extracts x- and y-coordinates along the border between the black and white fields in the images returned from Mechanical Turk workers.

A restriction of the “import_jpg” function is in how an image’s points are sequenced the same from tooth to tooth. So, for instance, the first point listed for one tooth could correspond to the extreme left of the image and the first point listed for a different tooth could be the point on the extreme right. This creates problems when using landmark based approaches such as calculating distances between Mechanical Turk workers and the gold standard teeth extracted by an expert. In order to overcome this, we took the points extracted from “import_jpg” and first performed EFA using the function “efourier” on the (x, y)-coordinates to derive harmonics that describe the tooth.

Elliptical Fourier Analysis is specified as a parametric function

$$x = f(t) = A_0 + \sum_{j=1}^H a_j \cos(jt) + \sum_{j=1}^H b_j \sin(jt)$$

$$y = f(t) = C_0 + \sum_{j=1}^H c_j \cos(jt) + \sum_{j=1}^H d_j \sin(jt)$$



Fig 3. Three black and white images produced by Mechanical Turk workers for the tooth shown in Fig 1.

<https://doi.org/10.1371/journal.pone.0179757.g003>

where H is the number of harmonics used, A_0 and C_0 are constants, and a_h , b_h , c_h , and d_h are the amplitudes associated with the h -th harmonic and $h = 1, 2, \dots, H$. Since EFA is not a landmark based procedure, the initial ordering of the points does not hinder the estimation of the harmonics. Next, so that we are able to perform landmark based analysis, we used the estimated harmonics to output a specific number of points around the edges of each tooth which all begin in the same location. These resulting points act as landmarks, which were used to calculate Riemann distance between shapes created by Mechanical Turk workers and created by the expert.

Additionally, the amplitudes (i.e. a_h , b_h , c_h , and d_h) created in EFA can then be used as input features in machine learning algorithms to classify the teeth to tribes and species. Since ultimately what we are interested in is classifying these teeth, the performance of classifiers based on the work of Mechanical Turk workers was compared to the classification accuracy when the model was trained using the outlines traced by the expert. The classification algorithm considered here was random forests [28]. The tracings from the Mechanical Turk worker and the expert were compared to assess how similar they are and to assess differences in the predictive accuracy.

In order to measure the tracing error, Riemannian distance [21] was calculated between the Turkers tracings and the expert tracing. To do this, we first extracted the edges of the black and white images using the “import_jpg” function in the “Momocs” [19] package in R. This creates a given number of (x, y)-coordinates for the outlines of the black and white images. However, the ordering of these points may not line up correctly with the ordering of another tracing of the same tooth. These harmonics can then be used as input in the function “efourier_shape” to output 150 (x, y)-coordinates which act as landmarks around each tooth so that a direct comparison can be made between the mechanical Turk tracings and the tracings performed by the expert.

3.3 Evaluation of Mechanical Turk work

In order to measure the tracing error, Riemannian distance [29] was calculated between the tracings generated by Turkers and the expert tracing. To do this, we first extracted the edges of the black and white images using the “import_jpg” function in the “Momocs” [27] package in R [30]. This creates a given number of (x, y)-coordinates for the outlines of the black and white images. However, the ordering of these points may not line up correctly with the ordering of another tracing of the same tooth. These harmonics can then be used as input in the function “efourier_shape” to output 150 (x, y)-coordinates which act as landmarks around each tooth so that a direct comparison can be made between the mechanical Turk tracings and the tracings performed by the expert.

Ultimately the goal of tracing these outlines is to accurately classify the tribe and species that these teeth represent. Previous work [31] compared five different machine learning algorithms based on their performance classifying teeth into tribe and species. Here, we only consider the use of random forests for classification of tribe to compare the tracings created by mechanical Turk workers to the tracings created by JKB.

4 Results

4.1 Tracing error

The Riemannian error distances ranged from 0.01113 to 1.113 with a median error of 0.1154. A histogram of this distribution can be seen in Fig 4. Notice that the distribution is skewed heavily to the right and indicates that many of the Mechanical Turk workers trace the outline with only small amounts of error with a full 50% less than 0.1154. For reference, Figs 5 and 6 show two examples of the work of Mechanical Turk workers, with outlines in red, yellow, and blue, compared to the gold standard, which is shown in black. In Fig 5, an example of the results for a tooth is shown. The Mechanical Turk tracings are visually nearly identical to the gold standard tracing and these correspond to Riemann distances of 0.0191, 0.0527, and 0.0342 for red, blue, and yellow, respectively. The other image in Fig 6 displays a different tooth where the Mechanical Turk workers struggled a bit more to accurately trace the outline of the occlusal surface relative to the gold standard. Visually the yellow tracing is the most accurate relative to the gold standard and has the lowest Riemann error of 0.0517. The tracings displayed by the red and blue curves are less accurate and correspond to Riemann errors of 0.1516 and 0.0879, respectively.

With a frame of reference for the meaning of the Riemann errors, Figs 7 and 8 display box-plots corresponding to the distribution of the Riemann distance by tooth position and Tribe. In Fig 7 it is evident that first lower molars (LM1) have much larger errors between the Mechanical Turk workers and the expert among tooth positions considered here.

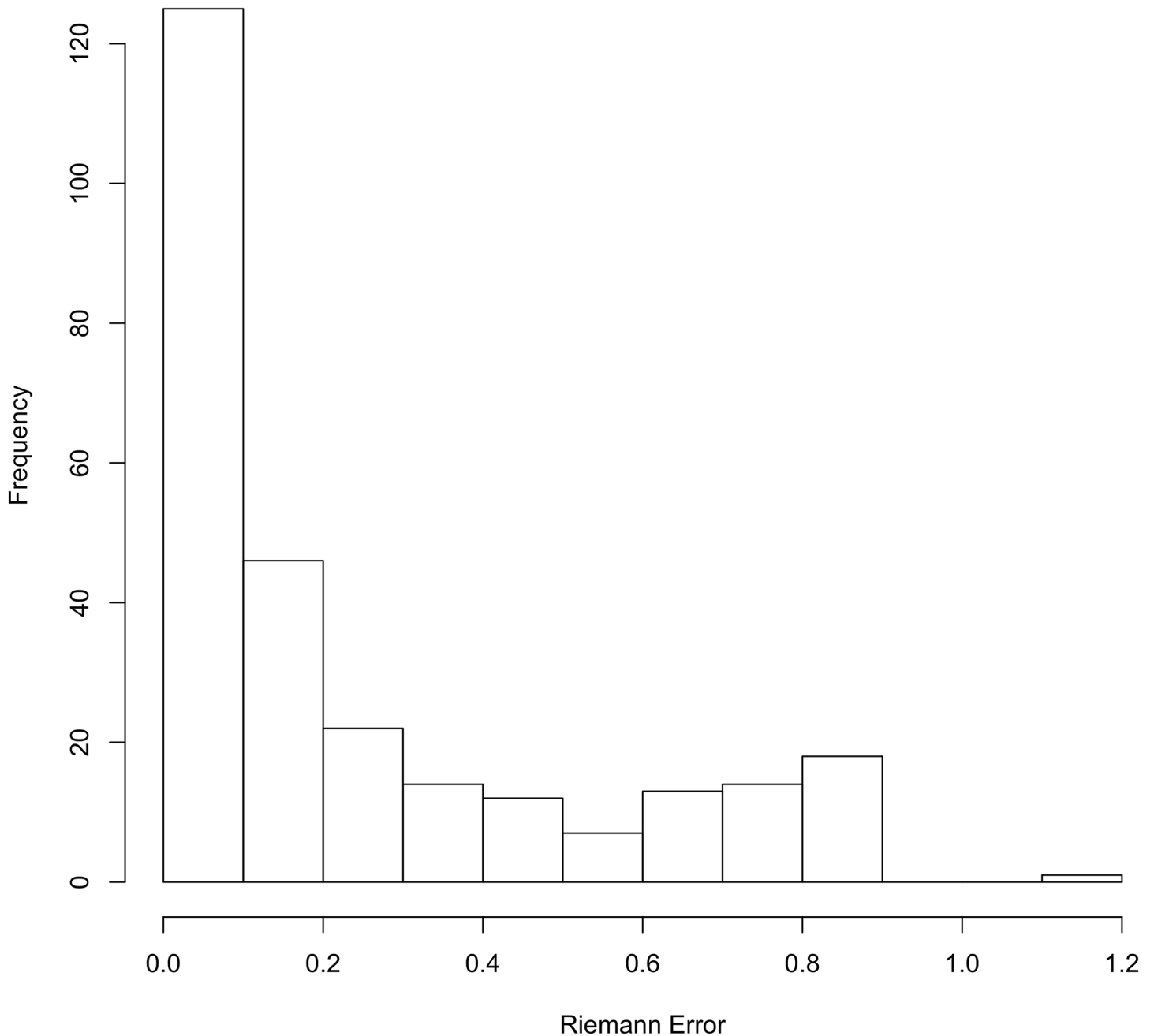


Fig 4. Distribution of errors as measured by Riemannian distance.

<https://doi.org/10.1371/journal.pone.0179757.g004>

4.2 Predictive accuracy

The histogram seen in Fig 9 depicts the classification accuracy results from the crowdsourced tracings. These results were created by repeatedly sampling one of the at-most three tracings per tooth in order to make a data set. Leave-one-out-cross validation was then performed using random forests. Accuracy of the model was quantified using a log loss score, comparing the predicted class to the actual observed class. From the histogram, it can be seen that if only

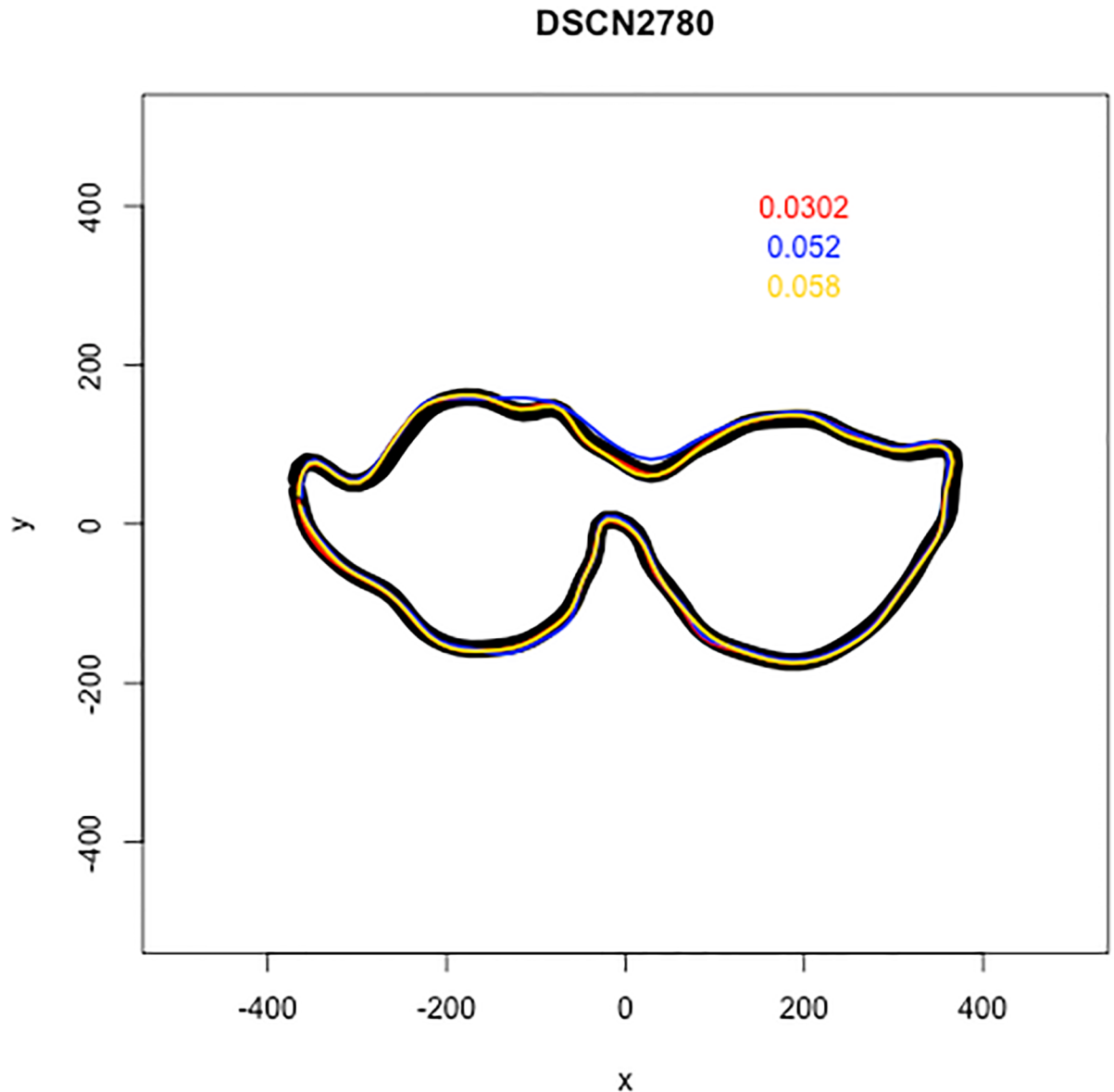


Fig 5. The black outline is the tracing done by JKB and the three other teeth in red, blue, and yellow correspond to the three tracings done by the Mechanical Turk workers. The red, blue, and yellow numbers that appear the upper right of the image correspond to the Riemannian distance between each Mechanical Turk tracing and the tracing done by the expert.

<https://doi.org/10.1371/journal.pone.0179757.g005>

one Turker for each image was used, they would perform consistently and considerably worse than the expert. The best sample is roughly .85 in terms of log-loss, while the mean is closer to 1.3, while the worst case is nearly 1.5.

The dotted line labeled “Mean of MTurk” was calculated by classifying the average shape of the Turkers outlines after eliminating obviously incorrect tracings. One can see that there is an improvement over even the best sample of individual workers. By taking the average image, the log-loss value lowered to 0.7788 for classifying the tribe.

Using the expert’s tracings we can further reduce log-loss, which is to be expected, down to 0.6689. While this is certainly an improvement over the Mechanical Turk workers, we argue

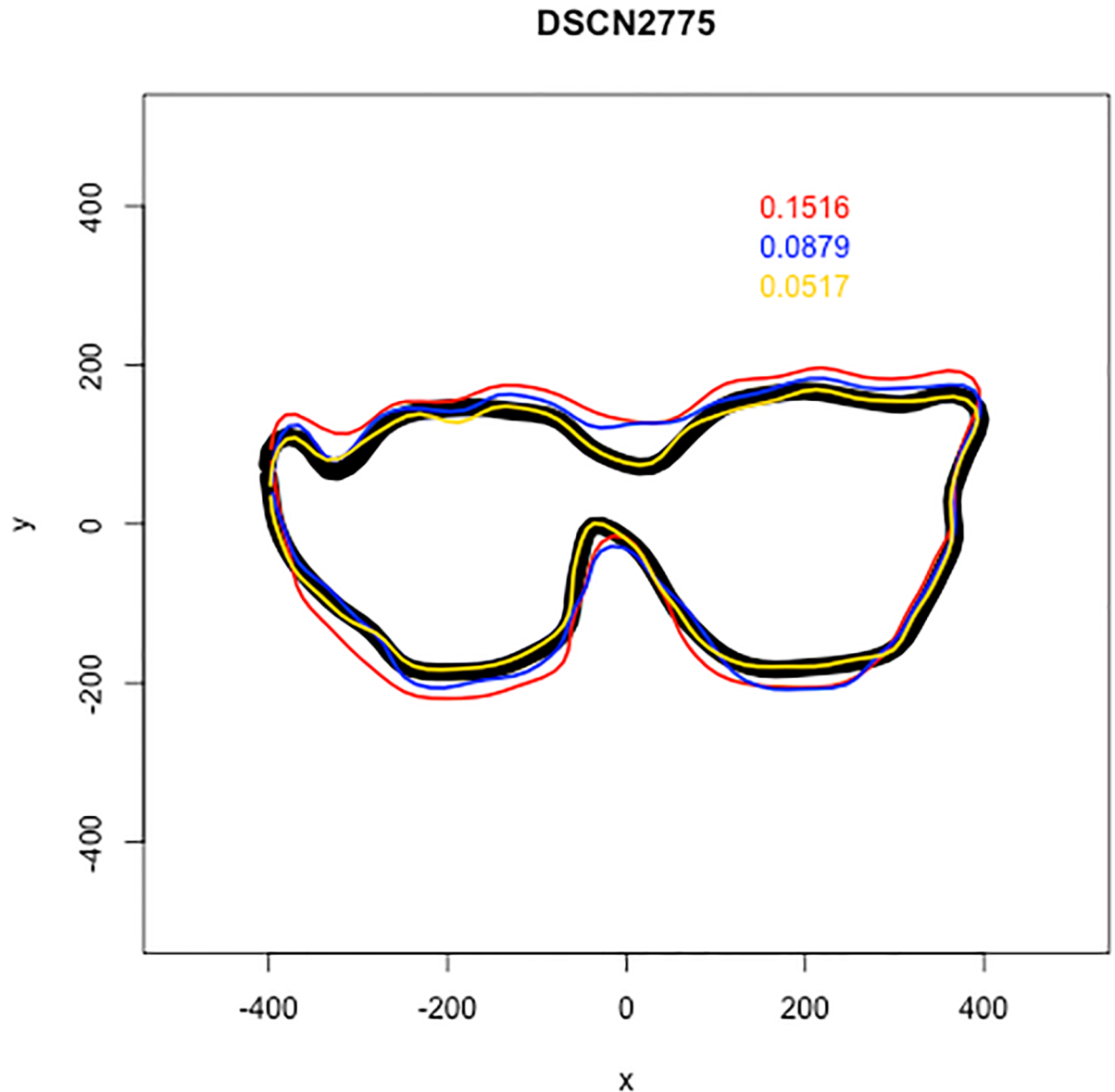


Fig 6. The black outline is the tracing done by JKB and the three other teeth in red, blue, and yellow correspond to the three tracings done by the Mechanical Turk workers. The red, blue, and yellow numbers that appear the upper right of the image correspond to the Riemannian distance between each Mechanical Turk tracing and the tracing done by the expert.

<https://doi.org/10.1371/journal.pone.0179757.g006>

that this level of log-loss is still acceptably close to the expert to still be of use in that the time that is saved by crowdsourcing the extraction of the edges is worth a small trade-off in classification accuracy.

Finally, we evaluated the classification performance of the traced outlines by averaging all of the Mechanical Turk workers (excluding images where the Riemann distance was greater than 0.2 from the expert) and the expert. This slightly improved classification accuracy compared to the average of the Mechanical Turk workers to a log-loss of 0.7524; however, the expert alone still has the lowest log-loss.

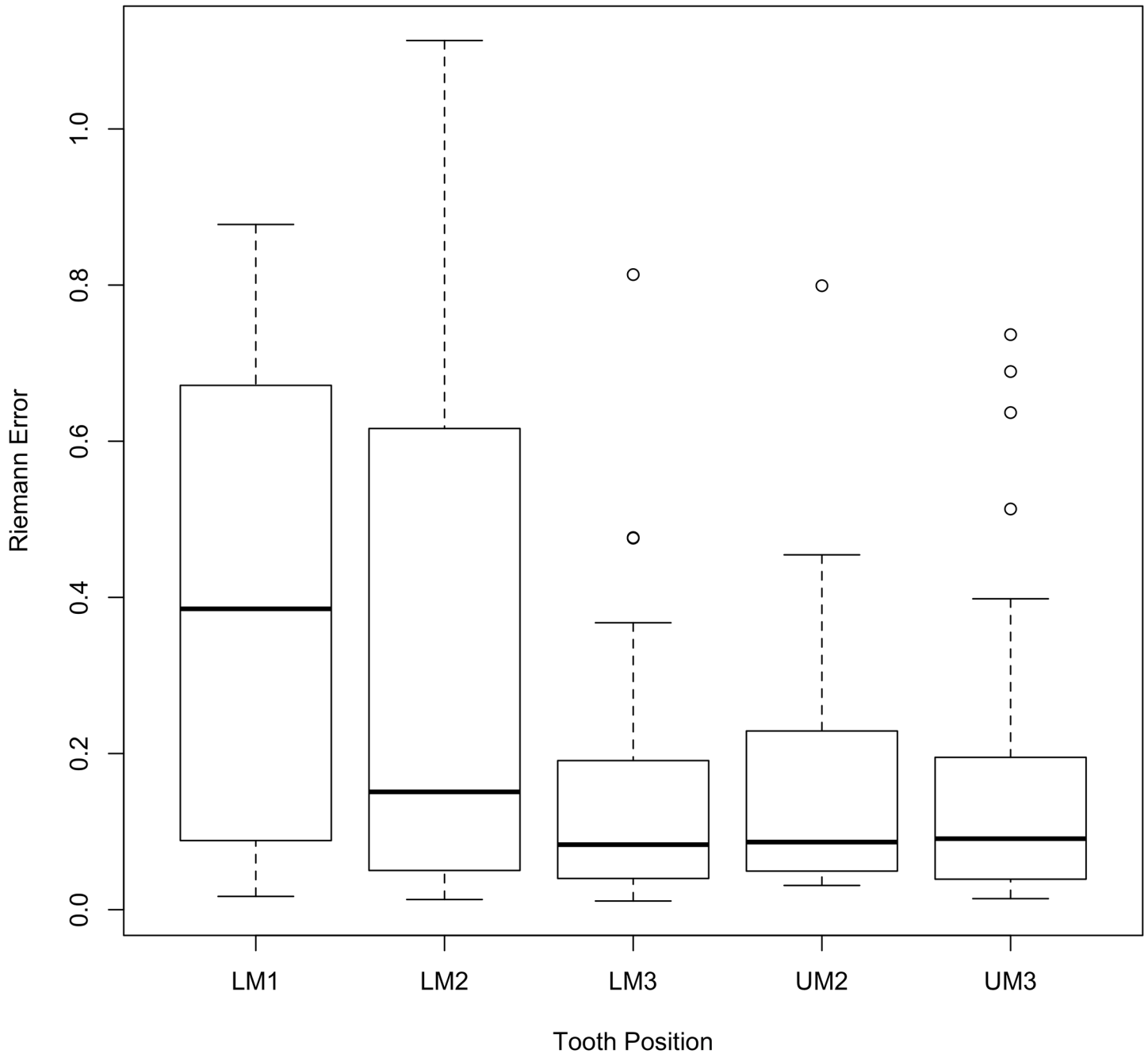


Fig 7. Errors as measured by Riemannian distance by tooth position.

<https://doi.org/10.1371/journal.pone.0179757.g007>

Finally, we consider results in terms of misclassification rather than log-loss. Table 3 shows the misclassifications for JKB alone. Using only those tracings, the model was able to classify correctly 79% of the specimens in cross validation. A large amount of the error occurred between Alcelaphini and Hippotragini. Namely, of the misclassified observations, 75% were either actually Alcelaphini but classified as Hippotragini, or actually Hippotragini but classified as Alcelaphini. Table 4 shows the misclassification results of the average image from the

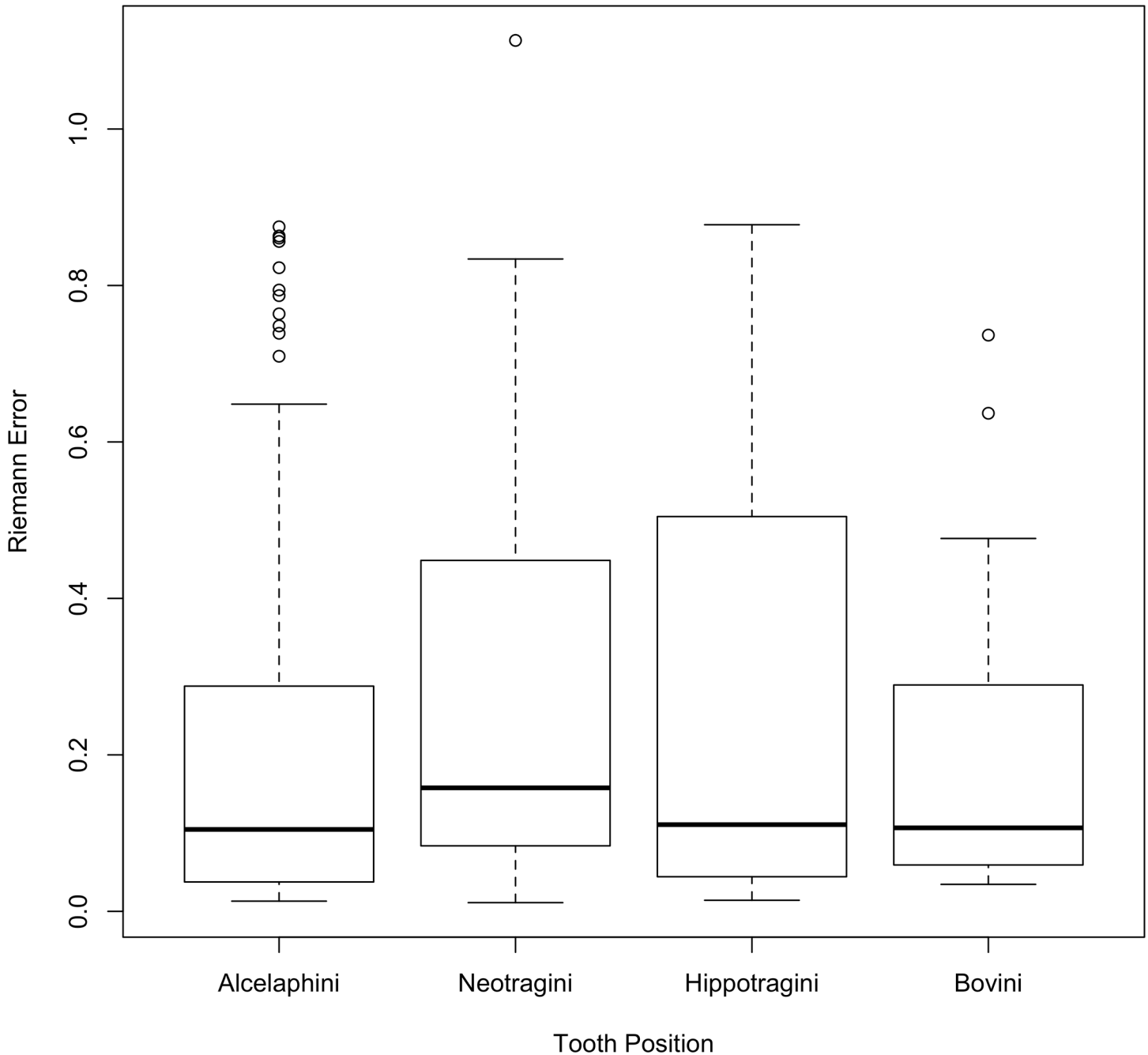


Fig 8. Errors as measured by Riemannian distance by Tribe.

<https://doi.org/10.1371/journal.pone.0179757.g008>

Turkers. The model correctly classified the Turker results 74% of the time. Once again, the largest source of confusion was between Hippotragini and Alcelaphini. Table 5 shows the results when the outlines of the Mechanical Turk workers were averaged with the gold standard. Somewhat surprisingly, this result was worse in terms of missclassification than the other two specifications considered here with a classification rate of 68% in spite of being better than using the Turk outlines only in terms of log-loss.

Distribution of Log-Loss Scores

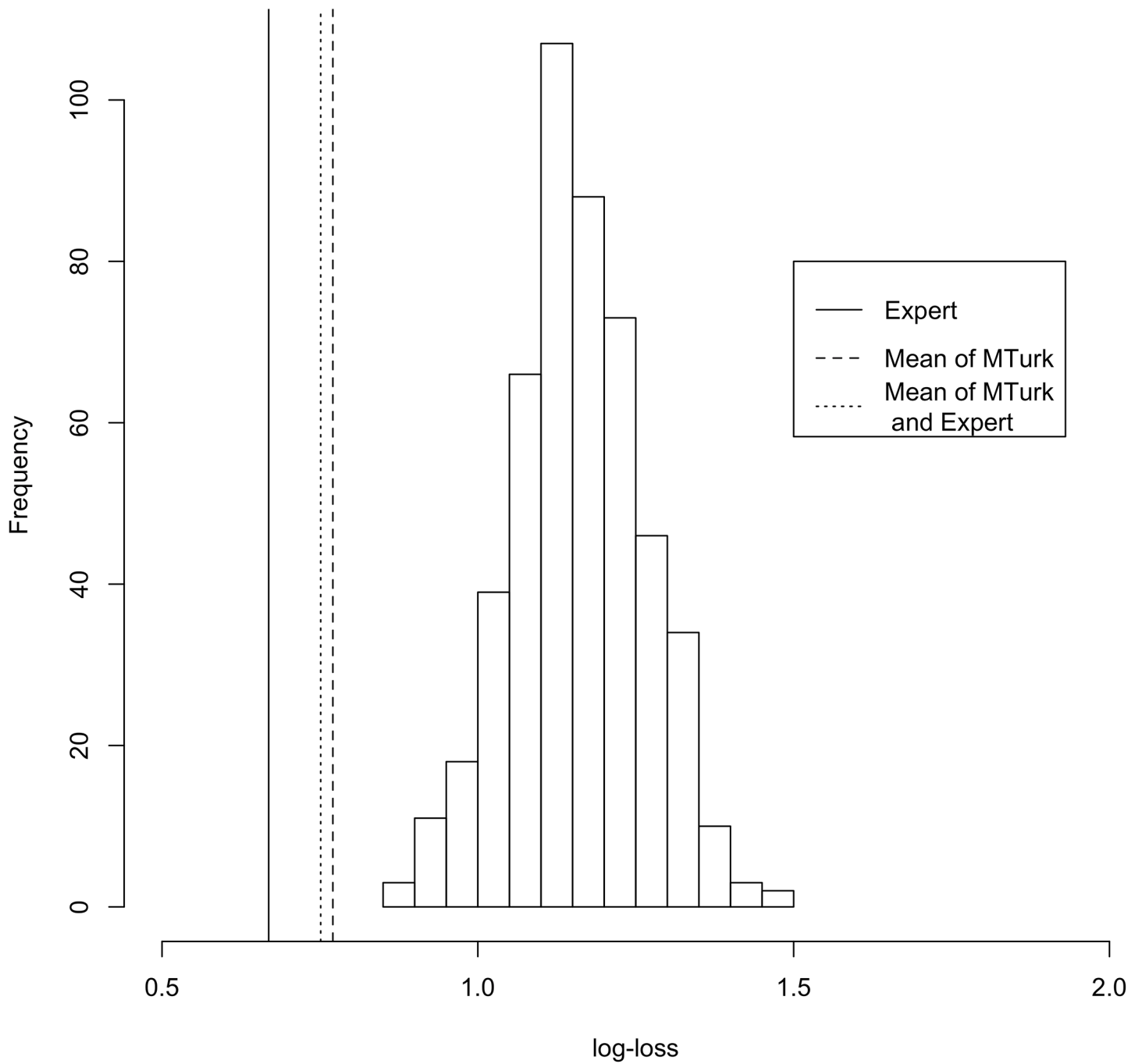


Fig 9. Comparing the classification accuracy of different methods of extracting the edges of bovid teeth.

<https://doi.org/10.1371/journal.pone.0179757.g009>

Table 3. Expert tracings only.

Actual Class	Predicted Class		
	Alcelaphini	Hippotragini	Neotragini
Alcelaphini	11	3	1
Hippotragini	3	11	1
Neotragini	0	0	8
Log Loss: 0.6688741			

<https://doi.org/10.1371/journal.pone.0179757.t003>

Table 4. Mean of MTurk worker tracings (Removed if Riemann distance from mean > 0.2).

Actual Class	Predicted Class		
	Alcelaphini	Hippotragini	Neotragini
Alcelaphini	10	2	3
Hippotragini	4	11	0
Neotragini	1	0	7
Log Loss: 0.7787713			

<https://doi.org/10.1371/journal.pone.0179757.t004>

Table 5. Mean of MTurk workers and expert tracings.

Actual Class	Predicted Class		
	Alcelaphini	Hippotragini	Neotragini
Alcelaphini	11	3	1
Hippotragini	5	8	2
Neotragini	1	0	7
Log Loss: 0.7524494			

<https://doi.org/10.1371/journal.pone.0179757.t005>

5 Discussion

The results of this study suggest that the proposed method will dramatically decrease the amount of subjectivity in bovid tooth identification and will advance the field of paleoanthropology/zooarchaeology. The importance of this method cannot be understated. As mentioned previously, bovinds have different ecological requirements. Therefore, misidentified bovinds can lead to incorrect paleoenvironmental reconstructions. For example, three researchers analyzed the bovid fauna from the South African site of Makapansgat and proposed paleoenvironmental reconstructions for Member 3 [32–34]. While each researcher relied upon the same assemblage to form their reconstruction, the papers suggest a different paleoenvironment: shrub-like with nearby open grasslands [32]; woodland [33]; and bushland with riparian woodland and nearby limited wetlands [34]. Reconstructions like these are used to discuss hominin behavior as well as speciation and extinction events. In fact, until recently it was commonly thought that one early human ancestor, *Australopithecus robustus*, went extinct due to being a habitat specialist that could not survive in fluctuating environmental conditions [4]. By more accurately identifying the bovinds from sites associated with *A. robustus* using morphometrics, [4] was able to demonstrate that this hominin lived in a variety of habitats that changed over time; *A.*

robustus was more likely a habitat generalist. Therefore, the hypothesis that *A. robustus* went extinct because it was a habitat specialist requires rethinking. If a fraction of these subjectivity problems are solved with this new methodology, the field is advancing and more accurate paleoenvironmental reconstructions and interpretations will be made.

With that said, some preliminary issues exist with this methodology. First, if a large number of teeth needs to be traced with replicates of each tooth, this process can get expensive. In the future, ideally, we will be able to leverage modern computer vision algorithms to extract the edges of these teeth with little or possibly no human aid. Second, some teeth are more difficult for a lay person to trace (e.g. LM1) and those teeth may still require an expert to trace those teeth or at least someone who has received more training than the average Mechanical Turk worker. This result is not unexpected as this method is not designed to completely replace all other forms of tooth identification, rather it is intended to provide objective, reliable classifications of bovid teeth and to supplement and be supplemented by other forms of tooth identification, as needed. Regardless of these problems, the benefits of employing this method and decreasing the subjectively involved in bovid tooth identification far outweigh the issues.

6 Conclusion

This study demonstrates that by taking the average shape of multiple Mechanical Turk workers, we can quickly obtain the outline the occlusal surface of a tooth that performs similarly to the expert's in terms of classification. A database was created of 96 different teeth along with the associated ground truth tracings done by an expert. Once outlines traced by non-experts through Amazon's Mechanical Turk were collected, we imported an outline into R and lined up landmarks for comparison using EFA. The accuracy of the tracings was evaluated by calculating the Riemann distances between the landmarks on the crowdsourced outline and the outlines generated by the expert. Further, predictive accuracy was assessed using leave-one-out cross validation with random forests on a small subset of the data. We find that in terms of log-loss the tracings performed by the expert, while superior, were not substantially better than using the average of the mechanical Turk workers. In terms of classification accuracy, we measured 74% classification rate using the average of the tracings of the mechanical Turk workers, which is very close to the classification accuracy of 79% when using the tracings generated by the expert. The results suggest that this process can be useful for researchers in many scientific areas (e.g. anthropologists, paleontologists, zooarchaeologists, etc.) who need quick, objective classifications for teeth recovered in the field. Further, one area of future work we are particularly interested in is the analysis and classification of partially observed teeth due (i.e. broken teeth). We believe that this method explored here can be easily extended to the case when teeth are broken.

Author Contributions

Conceptualization: GKT GJM JKB.

Data curation: GJM GKT MPL.

Formal analysis: GJM MPL.

Investigation: GKT GJM MPL.

Methodology: GJM GKT.

Software: GKT GJM MPL.

Visualization: GJM.

Writing – original draft: GJM JKB MPL GKT.

Writing – review & editing: GJM JKB MPL GKT.

References

1. Brophy JK, de Ruiter DJ, Athreya S, DeWitt TJ. Quantitative morphological analysis of bovid teeth and its implications for paleoenvironmental reconstructions in South Africa. *Journal of Archaeological Science*. 2014; 41(41):376–388. <https://doi.org/10.1016/j.jas.2013.08.005>
2. Bobe R, Eck GG. Responses to African bovids to Pliocene climatic change. *Paleobiology Memoirs*. 2001; 2:1–47. [https://doi.org/10.1666/0094-8373\(2001\)027%3C0001:ROABTP%3E2.0.CO;2](https://doi.org/10.1666/0094-8373(2001)027%3C0001:ROABTP%3E2.0.CO;2)
3. de Ruiter DJ, Brophy JK, Lewis PJ, Churchill SE, Berger LR. Faunal assemblage composition and paleoenvironment of Plovers Lake, a Middle Stone Age locality in Gauteng Province. *Journal of Human Evolution*. 2008; 55:1102–1117. <https://doi.org/10.1016/j.jhevol.2008.07.011> PMID: 18954892
4. Brophy JK. Reconstructing the habitat mosaic associated with *Australopithecus robustus*: evidence from quantitative morphological analysis of bovid teeth. Texas A&M University; 2011.
5. Adams DC, Rohlf FJ, Slice DE. The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments. *Hystrix*. 2013; 24:7–14.
6. Bright JA. A review of paleontological finite element models and their validity. *Journal of Paleontology*. 2014; 88:760–769. <https://doi.org/10.1666/13-090>
7. Drake AG, Coquerelle M, Colombeau G. 3D Morphometric analysis of fossil canid skulls contradicts the suggested domestication of dogs during the late Paleolithic. *Scientific Reports*. 2015; 5:8299. <https://doi.org/10.1038/srep08299> PMID: 25654325
8. Falkingham PL. Acquisition of high resolution three-dimensional models using free, open-source, photogrammetric software. *Palaeontologia Electronica*. 2012; 15(1):15p.
9. Klingenberg C. Evolution and development of shape: integrating quantitative approaches. *Nature Reviews Genetics*. 2010; 11:623–635. PMID: 20697423
10. Mitteroecker P, Gunz P. Advances in Geometric Morphometrics. *Evolutionary Biology*. 2009; 36:235–247. <https://doi.org/10.1007/s11692-009-9055-x>
11. Polly PD, Stayton CT, Dumont ER, Pierce SE, Rayfield EJ, Angielczyk KD. Combining geometric morphometrics and finite element analysis with evolutionary modeling: towards a synthesis. *Journal of Vertebrate Paleontology*. 2016; <https://doi.org/10.1080/02724634.2016.1111225>
12. Kuhl FP, Giardina CR. Elliptic Fourier features of a closed contour. *Computer graphics and image processing*. 1982; 18(3):236–258. [https://doi.org/10.1016/0146-664X\(82\)90034-X](https://doi.org/10.1016/0146-664X(82)90034-X)
13. Canny J. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence*. 1986; 8:679–714. <https://doi.org/10.1109/TPAMI.1986.4767851>
14. Lindeberg T. Edge Detection and Ridge Detection with Automatic Scale Selection. *International Journal of Computer Vision*. 1998; 30(2):117–154. <https://doi.org/10.1023/A:1008097225773>
15. Wolfe CA, Lestrel PE, Read DW. EFF23 2-D and 3-D Elliptical Fourier Functions; 1999.
16. Lestrel PE. Method for analyzing complex two-dimensional forms: Elliptical Fourier Functions. *Am J Hum Biol*. 1989; 1:149–164. <https://doi.org/10.1002/ajhb.1310010204> PMID: 28514011
17. Amazon's Mechanical Turk;. Available from: <https://www.mturk.com/mturk/welcome>.
18. Huff C, Tingley D. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk Survey Respondents. *Research & Politics*. 2015; 2(3).
19. Rouse SV. A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*. 2015; 43:304–307. <https://doi.org/10.1016/j.chb.2014.11.004>
20. Rand DG. The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments. *Journal of Theoretical Biology*. 2012; 299:172–179. <https://doi.org/10.1016/j.jtbi.2011.03.004> PMID: 21402081
21. Sprouse J. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*. 2011; 43:155–167. <https://doi.org/10.3758/s13428-010-0039-7> PMID: 21287108
22. Vittayakorn S, Hayes J. Quality Assessment for Crowdsourced Object Annotations. In: *Proceeding of British Machine Vision Conference (BMVC)*; 2011.
23. Sorokin A, Forsyth D. Utility data annotation with Amazon Mechanical Turk. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08)*. IEEE; 2008. p. 1–8.

24. Amazon Mechanical Turk, General Questions, What is a HIT?; Accessed: May 22, 2017. Available from: <https://www.mturk.com/mturk/help?helpPage=overview>.
25. The GIMP Development Team. GNU Image Manipulation Program (GIMP); 2016. Available from: www.gimp.org.
26. ImageMagick; 2016. Available from: www.imagemagick.org.
27. Bonhomme V, Picq S, Gaucherel C, Claude J. Momocs: Outline Analysis Using R. *Journal of Statistical Software*. 2014; 56(13):1–24. <https://doi.org/10.18637/jss.v056.i13>
28. Breiman L. Random Forests. *Machine Learning*. 2001; 45:5–32.
29. Kendall DG. Shape manifolds, Procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society*. 1982; 16:81–121. <https://doi.org/10.1112/blms/16.2.81>
30. R Core Team. R: A Language and Environment for Statistical Computing; 2016. Available from: <https://www.R-project.org/>.
31. Matthews GJ, Brophy JK, Luetkemeier MP, Gu H, Thiruvathukal GK. A comparison of machine learning techniques for taxonomic classification of teeth from the Family Bovidae; 2016.
32. Cooke HBS, Wells LH. Fossil Bovidae from the Limeworks Quarry, Makapansgat, Potgietersrust. *Paleont Afr*. 1956;.
33. Vrba ES. The fossil record of African antelopes (Mammalia, Bovidae) in relation to human evolution and paleoclimate. 1995;.
34. Reed KE. Using large mammal communities to examine ecological and taxonomic structure and predict vegetation in extant and extinct assemblages. *Paleobiology*. 1998; 24(03):384–408.