Horse Recognition Using Inertial Measurement Units

Wouter Visser
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
w.a.visser@student.utwente.nl

ABSTRACT

In recent years, the use of equine data from Inertial Measurement Units(IMU's) has been going up in research as well as medicine. When this IMU data is used, it is often very useful to know from which horse it is. This paper shows a method that allows data from different horses to be separated, with the use of Artificial Neural Networks. This system was created with the use of a so-called Long Short-Term Memory Neural Network, also known as an LSTM. However, this method is reliant on the fact that there is data available from the horse that needs to be recognized. Because of this, this paper also proposes a method to recognize if data from a horse is not yet available, using a softmax probability baseline.

Keywords

Machine learning, Equine gait, Horse recognition, Inertial Measurement Unit, Neural Network

1. INTRODUCTION

The way a horse walks also referred to as its gait can be split into a few different categories, as first described by Milton Hildebrand[1]. Carefully studying the discrepancies in this gait, known as lameness, is integral to finding any problems a horse might have. Because of this, equine veterinarians spend around 20% of their time monitoring the gait of horses[2]. However, the monitoring of this gait done by humans is often not very reliable, as seen by the fact that veterinarians often disagree on whether a horse is lame [3, 4]. So instead, the use of objective monitoring using Inertial Measurement Units(IMU's) has been introduced[5], which has been shown to be more reliable[6]. These IMU's are sensors that combine accelerometers and gyroscopes to record movement in all three dimensions. The reliability gained by IMU's, in combination with the increase in the use of data analysis in general[7], resulted in the fact that the use of IMU's in equine research has also been climbing[8]. Research like this often requires large amounts of data from specific horses.

However, the gathering of this data might not be as straightforward as it seems. The data is often gathered from multiple horses at the same time. Keeping the data separated and knowing what data comes from which horse might therefore become an issue. So, it would be convenient if it would be possible to check from which horse this IMU data is, as this would make entering new information easier. On top of this, such a system would lend extra credibility to the health checks that are done using IMU data. However, if such a system were to be deployed, a special problem might arise if no data from a horse is available. In this situation, this data would be assigned to a different horse, the one that was given as output by the system. To counter this, such a system should also be equipped to see if it has not yet seen a horse.

These problems combined lead to the following research questions:

- **RQ1:** How reliable can a horse be recognized using IMU data?
  - When confusing horses, what is the chance that it guesses a horse of the same discipline?
- **RQ2:** Can a horse recognition system recognize whether it has not seen a horse before?

2. RELATED WORK

Quite some research has already been done with the use of equine IMU data and machine learning[9, 10, 11]. However, of these studies, the one by Serra Bragança et al.[9] is of the most use. This paper is about research into the classification of equine gait into the walk, trot, Lcanter, Rcanter, Pace, Paso fino, Trocha, and Tolt. What makes this paper the most applicable is the fact that it is also about a multi-class classification problem, in which the classes are relatively close together. In the paper, the researchers tried to classify the gait with two different methods. First, they used the raw IMU data as input for an Artificial Neural Network(ANN). This ANN relied mostly on a so-called Long Short-term memory layer, a system that uses not just the current sample it is looking at, but also learns from the previous few samples. This system worked, but in general, it was less reliable than the other system they used. In the second system, the writers first extracted features from the IMU data using an algorithm described in a different paper[12]. The extracted data was then used for a different ANN, consisting mostly of a more general-use Fully Connected layer. As the paper describing the feature extraction algorithm[12] is not publicly available, the best system to use from the horse activity classification paper would be the Long Short-term memory solution.

The concept of recognition of an individual based on gait does already exist as well. For instance, research has already been done to recognize people based on their gait[13]. The method not only allows people to be recognized by

35th Twente Student Conference on IT July 2nd, 2021, Enschede, The Netherlands.

the way they walk but also gives information about what part of the data was important at which point of the gait. The fact that a similar system for people exists shows that the concept of recognition using gait is not entirely novel, which is promising for our research, as it shows this method has successfully been tried before, although on people rather than horses.

3. METHODS

3.1 Data Collection

The data was collected from 51 different healthy sport horses in a test in which the horses were asked to perform a number of tasks. These tasks included multiple different gaits and directions of movement. Of these horses, six were dressage horses, 27 were eventing horses, nine were showjumping horses, and nine were endurance horses. The data was collected using IMU sensors from the system created by EquiMoves[14]. The sensors were attached to the pelvis, withers, head, and each of the four limbs. These IMU sensors have a sampling frequency of 200Hz, and consist of a low-g accelerometer with a range of $\pm$ 16g, a high-g accelerometer with a range of $\pm$ 200g, and a 3D gyroscope with a range of $\pm$ 2000dps. The data was labelled with the gait and the direction of movement.

3.2 Data Preparation

All data processing was done using Matlab 2021a. Of all the available data, only straight-line trot was considered for this study. The straight-line trot data segments were extracted and assigned to either the train or test data set in a way that resulted in a $\pm$ 25-75% split between test and training data. This split cannot be exact, as the duration of the segments of straight-line trot is not always the same. A check was made if at least one of such segments of a single horse was available for both the test and training data set. This caused some horses to be excluded from the final data set, as not enough data was available to both train and test reliably. The segments were then split up into smaller sections of 0.2, 0.5, 0.8, 1, 2, and 3 seconds long.

3.3 Data Analysis

The data analysis was done using Python 3 and the TensorFlow Keras libraries. The machine learning algorithm used was a Long Short-Term Memory(LSTM) neural network. This structure of the neural network was based on the system developed by Serra Bragança et al.[9], and consisted of two LSTM layers with a width of 500, each with ReLu activations, which were then followed by two Fully Connected layers. The first of these layers had a width of 40, and also had ReLu activation. The final layer also functioned as the output layer, and as such had a width equal to the number of horses that needed to be recognized and a SoftMax activation. The neural network is trained with 15 epochs, and a categorical cross-entropy loss function. The optimizer for the neural network is the Adam[15] system with a learning rate of 0.001.

3.4 Recognition Algorithm

The system used here is based on a paper by Hendrycks and Gimpel[16]. This paper shows that when using a softmax activation in the output layer, the result of this activation can be used to see if a sample is from a non-existing category. This activation result is also known as the prediction probability. For our purposes, this means that if a horse is not in the database that is used to train the neural network, it will have a relatively low prediction probability. This principle is used in the final recognition algorithm. In this algorithm, the result from the neural network is only accepted if the output probability of this result is higher than a certain baseline. If this is not the case, it will be rejected, and the prediction will be that the horse is not yet in the database.

3.5 Experiments

In this part of the Methods, the various experiments that have been run and their purposes will be explained. The results for these experiments can be found in the Results section. The meanings of these results are elaborated upon in the Discussion.

3.5.1 Accuracy Experiment

This experiment has two different purposes. The first goal is to find whether the system works and how accurate it is. The second goal is to find which segment length works best for this neural network. To determine this, the data is split up into the sections described in the Data Preparation section. Then, for every time duration, ten new neural networks are trained. Each of these neural networks is then tested on the testing split.

3.5.2 Discipline Experiment

The purpose of this test is to provide insight into how the neural network determines from which horse a specific segment is. It does so by looking at if the model is wrong, what is the chance that it instead picked a horse of the same discipline. In this test, the data is split into segments of one second, as this was the length that worked best in the Accuracy Experiment. Ten models are then trained on the training data, after which each model is presented with all the testing data and asked which horse it is. For each of these models, two counters are kept, a correct horse and a correct discipline counter. For every correct horse, both the correct horse and the correct discipline counter will be increased. For every case where the correct discipline but the wrong horse was predicted, only the correct discipline counter will be increased. The counters are later used to determine the resulting accuracies of the system.

3.5.3 Probability Baseline Experiment

Since the Hendrycks and Gimpel paper[16] does not give a formula to determine the baseline for out-of-distribution detection, we should determine our own formula. To do this, we should know some statistics of the out-of-distribution samples, like the average, max and, standard deviation of the prediction probability. To this end, the data is split into segments of one second. After this, the following is done ten times: At random, 10% of the horses are labelled as unseen, and their data is removed from the train and test data sets. After this, a model is trained, and for each of the models, the data from the unseen horses are passed through the model and recorded what the probability is with which the result is predicted.

3.5.4 Unseen Horses Experiment

To determine how well this system could be used to see if a horse is not yet in the data set, an experiment has to be run. In this experiment, the data is split into segments of one second, and these segments are assigned to the train and test group as described in the Data Preparation part. Then, the following steps are repeated ten times: The horses are randomly split into two groups, seen and unseen, with the train part of the seen group being used to train the model, similarly as in the Probability Baseline Experiment. After this, the test segments of the seen group, as well as all the unseen segments, are passed onto the Recognition Algorithm. As no concrete

Table 1. Average Accuracy per segment duration

Duration in seconds	Average Accuracy
0.2	94.6%
0.5	95.5%
0.8	93.4%
1.0	95.6%
2.0	92.4%
3.0	86.3%

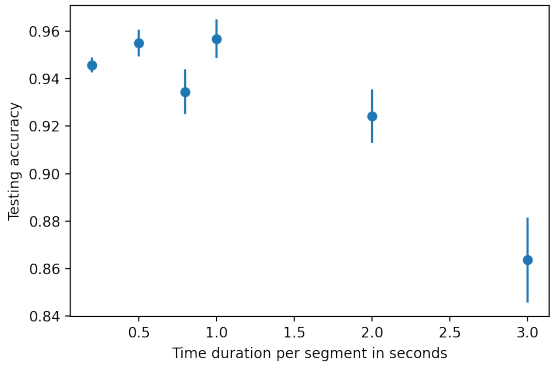

Figure 1. Average Accuracy per segment duration

formula is given by Hendrycks and Gimpel[16] to establish a baseline for the softmax function, a different formula was created. This formula is $MeanPredictionProbability + StandardDeviationPredictionProbability$

4. RESULTS

In this section, the results of the various experiments run are presented. An explanation of the experiments and their purposes is given in the Experiments part of the Methods. The impact of the results of the experiments is presented in the Discussion.

4.1 Accuracy Experiment

The average accuracy for each segment length can be found in table 1.

These results, together with an error bar, are represented in figure 1.

4.2 Discipline Experiment

When run, this experiment shows that with a mean horse correctness of 96.0%, there is a mean discipline correctness of 99.1%. This means that if it has the wrong horse, it on average still has the right discipline 75.0% of the time.

4.3 Probability Baseline Experiment

The results of this experiment give a total Mean Prediction Probability of 0.714, with a standard deviation of 0.226. The maximum of the recorded prediction probabilities was 1.000.

4.4 Unseen Horses Experiment

First, the baseline of the experiment should be calculated based on the Probability Baseline Experiment. The formula given in the Experiments combined with the previous results gives a threshold of $0.714 + 0.226 = 0.94$. The combined results with these settings come out to a mean total accuracy of 84.2%, with an average seen accuracy of 85,9%, and average unseen accuracy of 80.9%. The individual test results, with the percentage of unseen samples in the total test group, are shown in figure 2. Why this

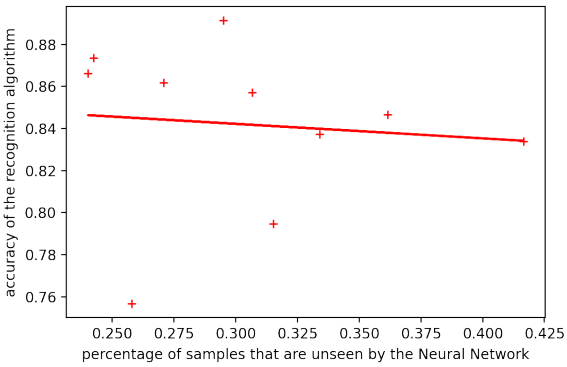

Figure 2. Average Recognition Accuracy

is important, and what can be learned from this graph, is explained in the Discussion.

5. DISCUSSION

To thoroughly grasp the result of this research, first, all the experiments have to be discussed one by one. This starts with the first experiment, the Accuracy Experiment. This experiment shows two different things. Firstly, it shows that this method could work to recognize horses based on IMU data. In the situation where one second of data was given, it has an average accuracy of 95.6%, which in most cases should be high enough to reliably recognize horses. Secondly, it shows that 1 second of data creates the best average accuracy of all the durations tested. This might have something to do with stride timings. Because on average, a trot has 91.3 ± 4.8 strides per minute[17]. This comes out to between 0.62 and 0.69 seconds per stride. If a time of 1 second is then taken, it is certain that at least one stride is available, with enough context around the stride. The 0.8-second segment length might not work as well because of the same reason, the fact that not enough context is available to properly predict which horse created the data segment. Another interesting point to note about the results of this experiment is the fact that in contrast to the findings of Serra Bragança et al.[9], 2 and 3 seconds work quite a bit worse than 1 second. The reason for this might have something to do with the fact that not a lot of data was available. Because of this, there are not a lot of 2 or 3-second segments available to learn from.

The Discipline Experiment gives some small insights into the way the recognition algorithm works. Because when it had the wrong horse, it still very often had the correct discipline. And because the algorithm was not trained with this information, we know that the things that differentiate the walking styles of different disciplines also can be used to recognize horses.

However, as there was no formula included in the original paper about the probability baseline[16], there was no definitive way to determine this baseline. But, there were still some other insights gained from this. Most notably, it showed that the maximum prediction probability was 1.0, meaning that some out-of-distribution samples had a prediction probability of 1.0. This is useful to show that it is not possible for the recognition system to entirely filter out all samples from unseen horses, as some wrong predictions will have a too high probability.

As for the final experiment, the Unseen Horses Experiment, this also gave some very interesting insights. First

of all, it shows that this system is both not filtering out all unseen samples, and does not keep all seen samples. However, it does still have a rather high accuracy, and it does not lower the accuracy of the seen horses to an unusable degree. This test also shows something else, that is important to note about the usage of this system. That has to do with the trend line of the accuracy against the percentage of unseen samples. This trend line shows that there exists a difference between the unseen accuracy and the seen accuracy. This is important to note because that shows that the baseline that was chosen prioritizes seen accuracy over filtering out all unseen samples. The graph and the statistics combined show that the baseline value that is used should be changed with the use case. If it is critical that most of the unseen horses are filtered out, this value can be set lower, and if unseen horses are unlikely to happen, this value can be set rather high.

An important thing to mention with the results of this system is that most likely such a system would not be run on one second of data. This is because when data is collected, it is often more than one second. So, the recognition algorithm will be run on all of this collected data, because this gives the highest accuracy. The results of all of these recognitions will then be combined to provide the final recognition. With that, having an accuracy of 84.2%, as is the case for the recognition algorithm, is not a large problem, as when a segment of a few seconds is loaded in, the combined accuracy of these seconds is a lot larger than this.

Another caveat that is important to note is the amount and type of data used in this experiment. All of the data used in this research was collected on a single day. This might create a few different issues. Firstly, the fact that only one day was used means that there were less data in general. When using one second of data for each segment, each horse on average only had 52 seconds of data in the training set, with some having a lot less.

The second problem stems from the fact that all data comes from a single day. Because while the test and train data sets are all selected from different times of the day, the recognition algorithm still only learns from one day of data. As such, not much can be said now about how effective this system would be on more long term examples.

6. CONCLUSION

All in all, this paper demonstrates that it is possible for a neural network to recognize a horse based on IMU data. We also showed that the way horses are recognized using this data is in a way that makes them more likely to confuse these horses with horses of the same discipline. Next to this, a basic way to recognize if a horse is not yet in the data set was created. All of these findings can be of use when collecting data from multiple horses, or this could be used to validate whether data used for a medical test is from the correct horse.

7. FUTURE WORK

While plenty of insights were found with this research, there is plenty more that can be discovered about techniques like this. One of these things is looking at other machine learning approaches than an LSTM. For example, for the solution to the gait classification problem, Serra Bragança et al.[9] found that analysis using extracted features worked best. Another approach that is used in other research regarding IMU data is the use of a Convolutional Neural Network[18, 19], a different type of neural network that also works well to learn from a series of time data. This is another potential solution that could be looked at to solve the horse recognition problem. However, due to time constraints, it was not possible for us to investigate how well this would work.

Another topic that could potentially be looked at, as already mentioned in the Discussion, is the use of longer-term data. Because for the current solution only short term data was available, the potential for use over multiple weeks, months, or even years is left unexplored. So while it is now known that our horse recognition system works for data from one day, it is not yet certain how this would work on a long-term problem.

8. REFERENCES

[1] M. Hildebrand, "Symmetrical gaits of horses," *Science*, vol. 150, no. 3697, pp. 701–708, 1965.

[2] J. B. A. Loomans, P. W. T. Stolk, P. R. van Weeren, H. Vaarkamp, and A. Barneveld, "A survey of the workload and clinical skills in current equine practices in the netherlands," *Equine Veterinary Education*, vol. 19, no. 3, pp. 162–168, 2007.

[3] K. Keegan, E. Dent, D. Wilson, J. Janicek, J. Kramer, A. Lacarrubba, D. Walsh, M. Cassells, T. Esther, P. Schiltz, *et al.*, "Repeatability of subjective evaluation of lameness in horses," *Equine veterinary journal*, vol. 42, no. 2, pp. 92–97, 2010.

[4] M. Hewetson, R. M. Christley, I. D. Hunt, and L. C. Voute, "Investigations of the reliability of observational gait analysis for the assessment of lameness in horses," *Veterinary Record*, vol. 158, no. 25, pp. 852–858, 2006.

[5] T. Pfau, H. Boultbee, H. Davis, A. Walker, and M. Rhodin, "Agreement between two inertial sensor gait analysis systems for lameness examinations in horses," *Equine Veterinary Education*, vol. 28, no. 4, pp. 203–208, 2016.

[6] M. J. McCracken, J. Kramer, K. G. Keegan, M. Lopes, D. A. Wilson, S. K. Reed, A. LaCarrubba, and M. Rasch, "Comparison of an inertial sensor system of lameness quantification with subjective lameness evaluation," *Equine Veterinary Journal*, vol. 44, no. 6, pp. 652–656, 2012.

[7] A. Kulakli and V. Osmanaj, "Global research on big data in relation with artificial intelligence (a bibliometric study: 2008-2019)," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, p. 31, 02 2020.

[8] S. Egan, P. Brama, and D. McGrath, "Research trends in equine movement analysis, future opportunities and potential barriers in the digital age: A scoping review from 1978 to 2018," *Equine Veterinary Journal*, vol. 51, no. 6, pp. 813–824, 2019.

[9] F. Serra Bragança, S. Broomé, M. Rhodin, S. Björnsdóttir, V. Gunnarsson, J. Voskamp, E. Persson-Sjodin, W. Back, G. Lindgren, M. Novoa-Bravo, C. Roepstorff, B. van der Zwaag, P. Van Weeren, and E. Hernlund, "Improving gait classification in horses by using inertial measurement unit (imu) generated data and machine learning," *Scientific Reports*, vol. 10, no. 1, 2020. cited By 1.

[10] A. Schmutz, L. Cheze, J. Jacques, and P. Martin, "A method to estimate horse speed per stride from one imu with a machine learning method," *Sensors*, vol. 20, p. 518, 01 2020.

[11] H. Darbandi, F. Serra Bragança, B. J. van der Zwaag, J. Voskamp, A. I. Gmel, E. H. Haraldsdóttir,

and P. Havinga, "Using different combinations of body-mounted imu sensors to estimate speed of horses—a machine learning approach," *Sensors*, vol. 21, no. 3, 2021.

[12] F. Serra Bragança, J. Vernooij, P. René van Weeren, and W. Back, "Validation of distal limb mounted imu sensors for stride detection and locomotor quantification in warmblood horses at walk and trot," *Equine Veterinary Journal*, vol. 48, no. S49, p. 17, 2016.

[13] F. Horst, S. Lapuschkin, W. Samek, K.-R. Müller, and W. Schöllhorn, "Explaining the unique nature of individual gait patterns with deep learning," *Scientific Reports*, vol. 9, p. 2391, 02 2019.

[14] S. Bosch, F. Serra Bragança, M. Marin-Perianu, R. Marin-Perianu, B. J. Van der Zwaag, J. Voskamp, W. Back, R. Van Weeren, and P. Havinga, "Equimoves: A wireless networked inertial measurement system for objective examination of horse gait," *Sensors*, vol. 18, no. 3, 2018.

[15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[16] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[17] M. H. Ratzlaff, B. D. Grant, R. Rathgeber-Lawrence, and K. L. Kunka, "Stride rates of horses trotting and cantering on a treadmill," *Journal of Equine Veterinary Science*, vol. 15, no. 6, pp. 279–283, 1995.

[18] O. Dehzangi, M. Taherisadr, and R. ChangalVala, "Imu-based gait recognition using convolutional neural networks and multi-sensor fusion," *Sensors*, vol. 17, no. 12, 2017.

[19] S.-M. Lee, S. M. Yoon, and H. Cho, "Human activity recognition from accelerometer data using convolutional neural network," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 131–134, 2017.