TöltSense Validation

Dr Torben A Rees, TöltSense Limited

A study to assess the accuracy of the TS gait classification algorithm was conducted in the autumn / winter of 2021/22. Horses were ridden and filmed with the TS equipment in place creating a log file of gait labels with timestamps. A web application was created where experienced observers could watch the videos and click buttons to indicate the gait, creating a series of gait labels and timestamps. Taking the aggregated classification from several judges as the ground truth, a comparison was then made to establish the accuracy of TS. It was shown that TS achieves a gait classification accuracy of up to 99.7%.

In addition, the accuracy and precision of hoof-on and hoof-off measurement was investigated. Frame-by-frame analysis of 100 fps video of a single 5-gaited Icelandic horse was used to create a series of hoof-on/off events to serve as a benchmark. The hoof-on timings from TS were then aligned with this signal using cross correlation and for each event a difference was calculated. The resulting distribution was then subjected to analysis. It is shown that the accuracy of TS as a tool for measuring beat quality, duty factor, and suspension is at least as good as the best possible result from analysis of 30 fps video.

Part 1: Validation of Gait Classification

Introduction

TöltSense (TS) is a gait classification and performance monitoring system designed for use with Icelandic horses. It provides real-time feedback to the trainer using voice synthesis as well as post-session graphical and statistical information. The system comprises a network of four wireless inertial sensors worn on the lower limbs of the horse and an app on a mobile phone carried by the rider. The principle is to detect stride events (hoof on/off times) with enough accuracy and precision to reliably discriminate which of the 5 Icelandic gaits is being shown, and in addition, to provide analysis of the quality of the gait in terms of beat, balance, suspension, speed, and stride length. To achieve this in real-time, TS uses efficient

algorithms and data streaming strategies, together with aggregation and error correction based on domain-specific knowledge.[1]

The kind of information that TS can generate and supply to the rider would normally require an expert eye on the ground or an in-depth frame-by-frame analysis of video. Such circumstances inevitably incur costs and are restrictive in terms of environment, such as being confined to a riding hall, or needing good lighting, or fair weather. Moreover, the volume and utility of information gathered this way is limited. By contrast, TS can instantly analyse the rhythm of the horse's gait and channel this information back to the rider without any external assistance or environmental restrictions. This approach creates a short feedback loop, resulting in accelerated learning for the horse and rider.

Further, the data is logged and can be collated over many sessions, which opens numerous possibilities for incremental performance improvement and research. We will show that TS's gait classification is as good as any expert human observer. Specifically, the gait classification has been shown to have 99% agreement with the classifications of a panel of 4 qualified Icelandic sport judges. In addition, the measurement of inter-limb stride event timings, which forms the basis of gait quality evaluation, is at least as accurate as an idealised frame-by-frame analysis of 30 fps video.[2] TS therefore massively increases the accessibility of detailed and reliable gait performance metrics to the rider and trainer, compared to traditional methods.

A wealth of research has been conducted in recent years in the field of horse gait analysis using inertial sensors, most of which is geared towards lameness detection, remedial farriery, or cutting-edge research into equine gaits in general. Several commercial systems have arisen from this endeavour, costing thousands or even tens of thousands of pounds to purchase. TS is not as technically advanced as these systems and does not aspire to cutting edge veterinary research. Rather, the aim is to produce an affordable and accessible system that is accurate and fast enough to support the training needs of amateur and professional Icelandic horse enthusiasts.

[1] For commercial reasons, this paper will not go into details about how these mechanisms work.
[2] We are not suggesting that TS is a replacement for a qualified riding instructor or can somehow make up for any lack of experience in horse training. Rather, TS is a tool that both riders and trainers can use to enhance their existing ways of working.

Materials and methods

Eight Icelandic horses of varying levels of training and ability were ridden and filmed while wearing the TS equipment at a horse farm in the UK. Some of the sessions were captured during warm-up for an oval track competition in an indoor arena. The rest were captured during a training day on an oval track.

At the beginning of each session, we started filming by recording the press of the TS app's "START" button before continuing to film without interruption (see Figure 1). This button press initiates the creation of a log file of gait classifications and timestamps, and recording it provides an easy way to line up the video with the log. Each video was trimmed to start exactly when the 'START' button was pressed so that the times in the video would correspond to the times in the app log.

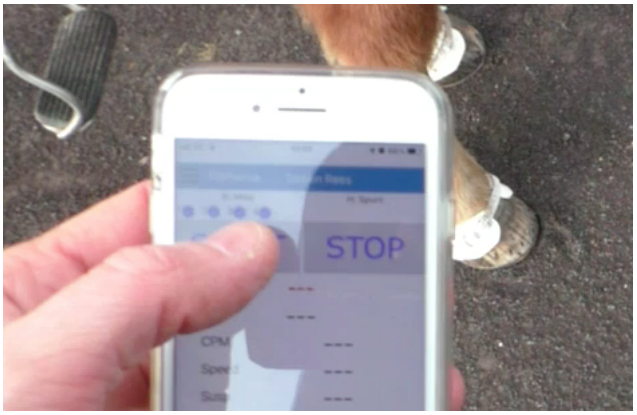

Figure 1: Filming the start event

A panel of 4 qualified Icelandic sport judges[3] independently suggested gait labels while watching the videos, using a custom-made web application (see Figure 2). For each video, a continuous observation window of 4-5 minutes was defined so as to include as many transitions and gaits as possible and to avoid judges having to watch unnecessary footage. At no point were the classifications of TS revealed to the judges.

[3] Ann Winter (IS, International Sport Judge), Nanco Lekkerkerker (NL, National Sport Judge), Rebecca Hughes (UK, Regional Sport Judge and IHSGB trustee), Harriet Vincent (UK, Regional Sport Judge and qualified Veterinary Surgeon)

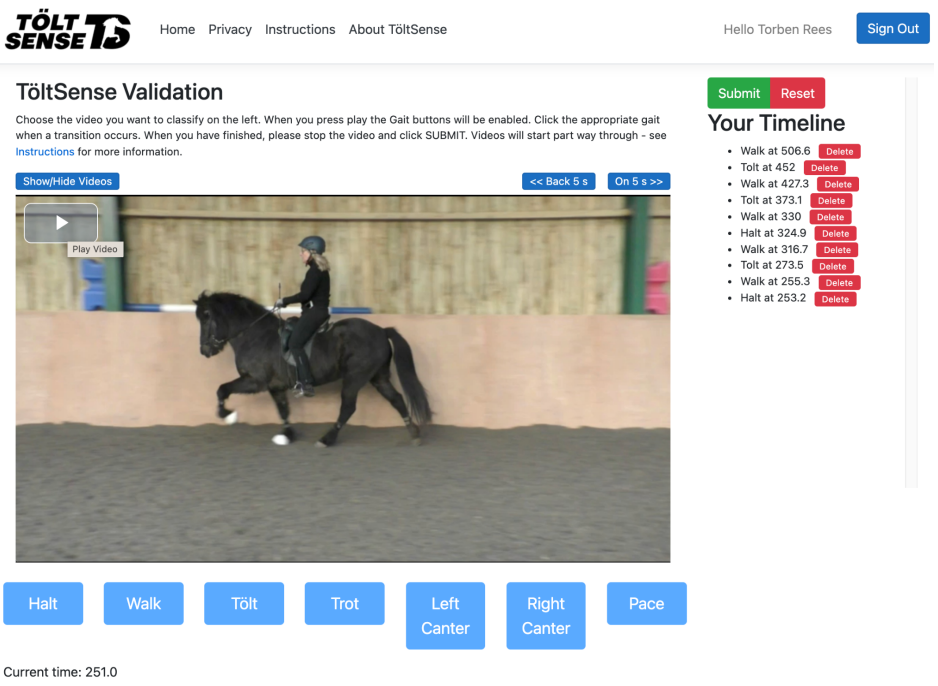

Figure 2: The validation web app

A gait was determined every 250 ms by choosing the most common label suggested during the given time interval. Such a majority vote was applied throughout the observation window, and a new data point was created whenever the majority gait changed. This processing step resulted in a series that defined the gait at any given moment during the observation window. We refer to this time series as *aggregate judge classifications*. An illustration of this time series is shown for a single example in Figure 3. The values in the time series correspond to gaits according to the following table:

No majority / disputed	-2
Not classified	-1
Halt	0
Walk	1
Trot	2
Tölt	3

Left Canter	4
Right Canter	5
Pace	6

The category “No majority / disputed” is required because there are occasions where judges do not agree on the gait. Mostly these are brief periods around gait transitions. However, Icelandic horses often display movements that are “between gaits” (e.g., pacey tölt, or 4-beated trot), and in these cases, it is not clear which gait is being shown. If qualified observers do not agree on the gait, there is no ground truth to compare TS against, so periods labelled as -2 were excluded from the assessment. “Not classified” appears once at the very start of each session before any classification has been given; such periods were also excluded. The other labels are self-explanatory but note that left and right canter are included as separate gaits because it is important to demonstrate that TS can distinguish between them.

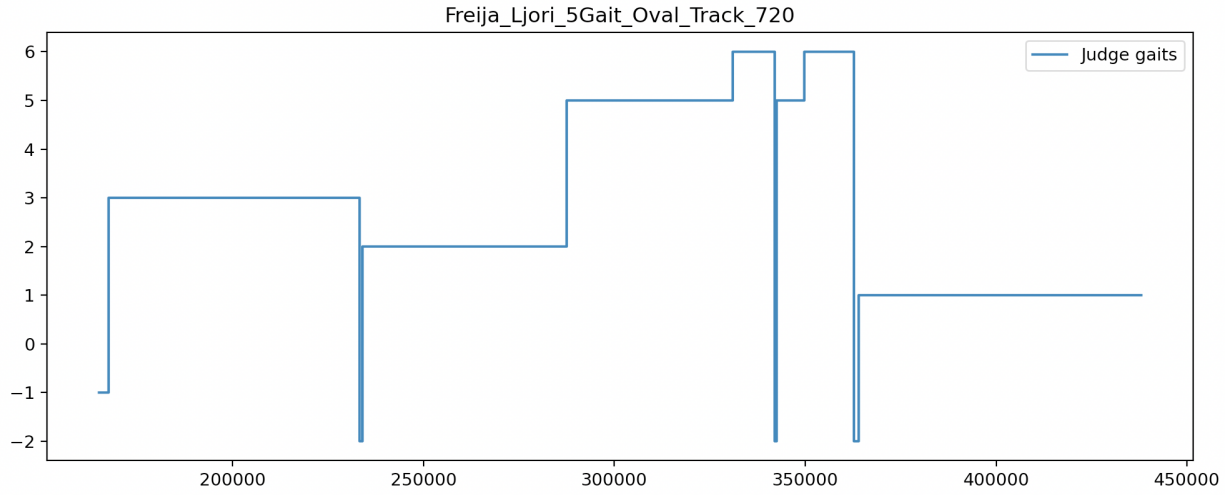

Figure 3: Plot of aggregate judge classifications against time, showing initial -1 period and 3 periods of -2 dispute

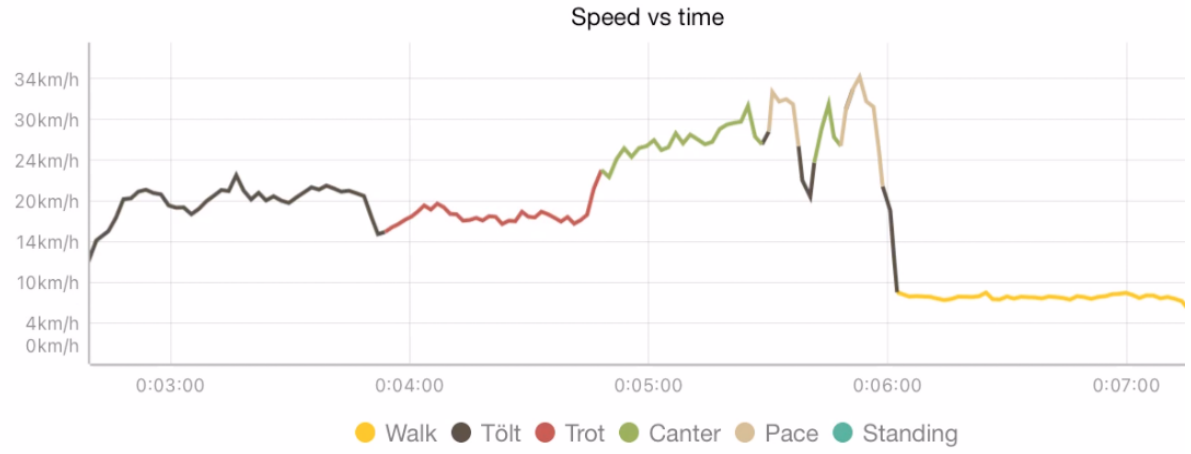

Figure 4: Plot of speed vs time colour coded by gait from TöltSense session overview screen

TS calculates the gait label every time a hoof-on event is registered from any leg. Hence the gait labels are produced at a variable rate from about 4Hz to 10Hz, depending on the horse's activity. We use a sliding window of 1s length with a step of 0.5s to produce a timeline for analysis. For each window, all gait labels falling within that window were collected and the most frequent gait label was taken and paired with the timestamp from the middle of the window. The timestamp of the first sample in the log was subtracted from all windows, which resulted in a timeline of gaits starting at 0s and progressing forward at 0.5 s intervals for the whole session.

At this point, we now have two synchronised series containing timestamps and gait labels. The series from TS should be regarded as the 'predicted value' set, and the series aggregated from the judges is the 'true value' set. For each video, an observation window has been defined where we have both a TS prediction and the ground truth labels generated by judges. We take the timestamp from the predicted value and retrieve the ground truth value at that given time to make a prediction-truth pair, or 'test case'.

As mentioned above, when aggregate judge classifications result in -2 or -1, the test case is discarded. In addition, it is appropriate to discard test cases that are very close to transitions. In other words, leeway should be allowed for both TS and judges when it comes to identifying and registering the moment of a transition. There are several reasons for this. First, there is always a delay between the horse making the transition, the judge recognising the transition has occurred, and then again a delay before the gait button is clicked in the web application. TS may already have correctly recorded the transition when the judge clicks

the button, so a false negative will be produced. Second, in cases of momentary loss or change of gait (for example, a few steps of trotty tölt in the middle of a section of trot going straight back to trot), a judge is unlikely to react fast enough to register the gait change. If they do, it is likely to be registered late.

On the other hand, TS is very likely to register the momentary gait change, and so again, a false positive or false negative may be produced. There are also cases where the human eye (and hand) is quicker to register the gait change than TS. The main example is the transition to walk, where the stride frequency is low, and it takes longer for TS to build a buffer of steps to analyse as opposed to when going into fast tölt, for example. In such cases, the judges may log a transition before TS does, which results in a false negative. For these reasons, we propose a 1s exclusion period for transitions (as identified by either TS or the judges) as a fair degree of leeway. Any test cases falling within these exclusion periods are removed from the analysis.

Results

A total of 4,421 one-second windows were generated from the TS logs, which included 179 gait transitions. Excluding cases that were subsequently marked as -2, or -1, and without exclusion periods, 4,371 valid test cases were generated. Factoring in transition exclusion periods reduces the number of valid test cases (see table below). The accuracy of the TS predictions was calculated simply as the total number of correct predictions divided by the total number of valid test cases.

The following table shows how the calculated accuracy varies when you alter the exclusion period for both judge-identified and TS-identified transitions.

TS Excl (ms)	**Judge Excl (ms)**	**Test Cases**	**Accuracy %**
2000	2000	3358	99.73
1000	1000	3757	98.86
1000	0	3909	97.95

0	1000	3990	96.62
0	0	4371	93.89

The exclusion period was varied from 0s to 2s, and it was found that the longer the period, the greater the accuracy. Without any exclusion periods, the accuracy was 93.89% and with 2s for each, the accuracy was 99.73%. The gait distribution is shown in Figure 5 and the confusion matrix for the case where both TS and Judge exclusion period was 1000ms is shown in Figure 6.

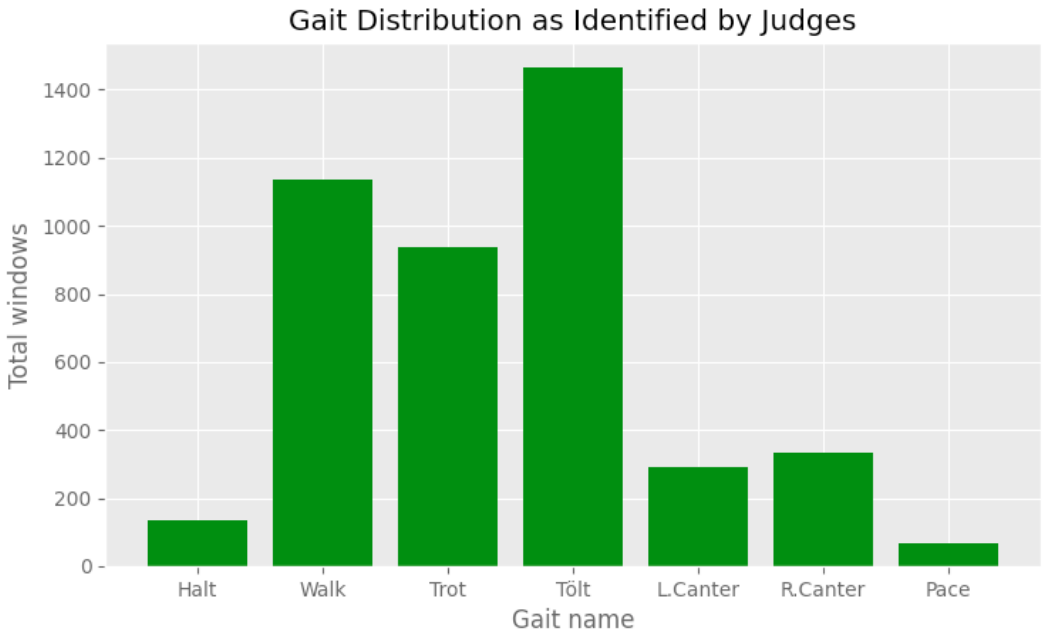

Figure 5: Distribution of gaits as identified by the judges was as below. We note that pace is underrepresented here.

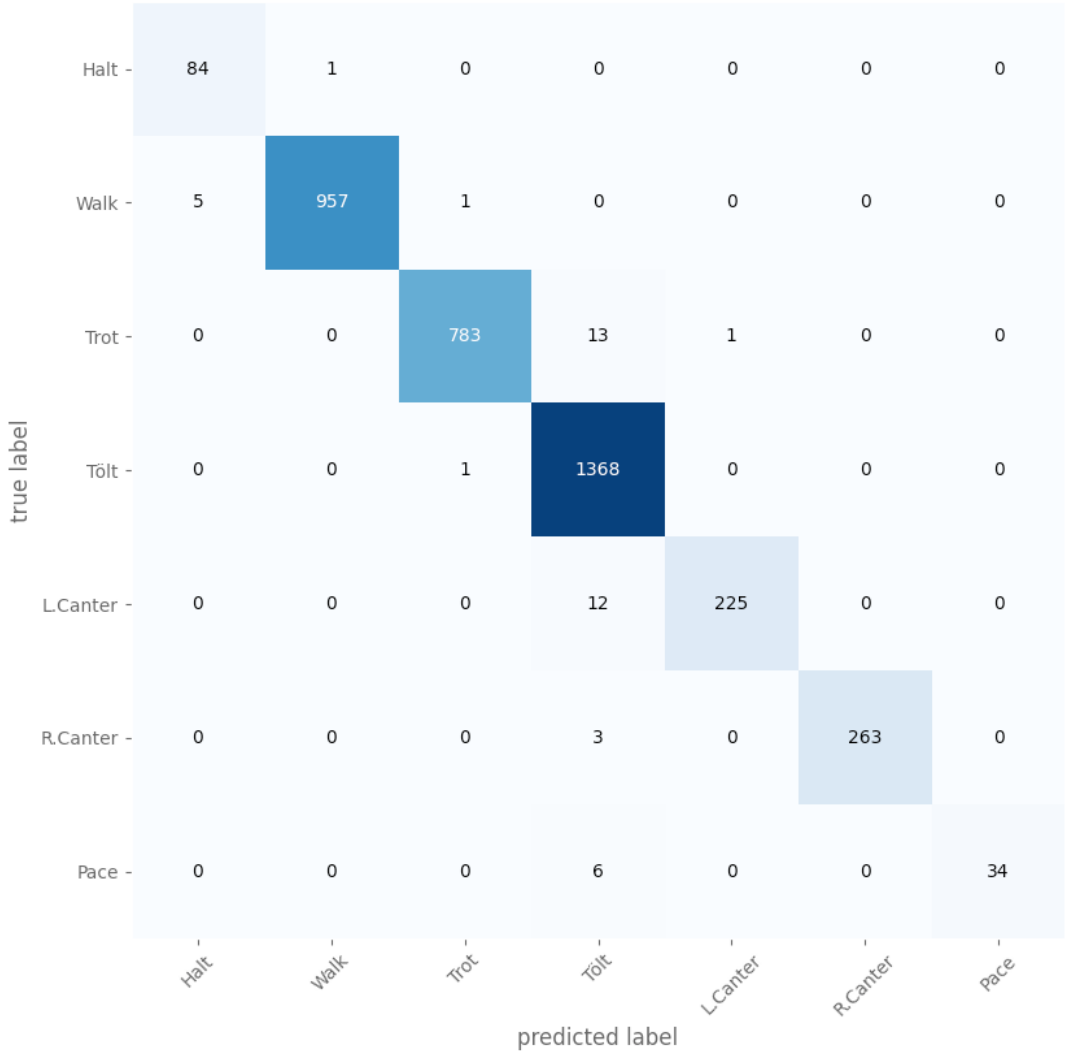

Figure 6: A confusion matrix for the case where both TS and Judge exclusion period was 1000ms.

The confusion matrix without any exclusion periods is shown in Figure 7.

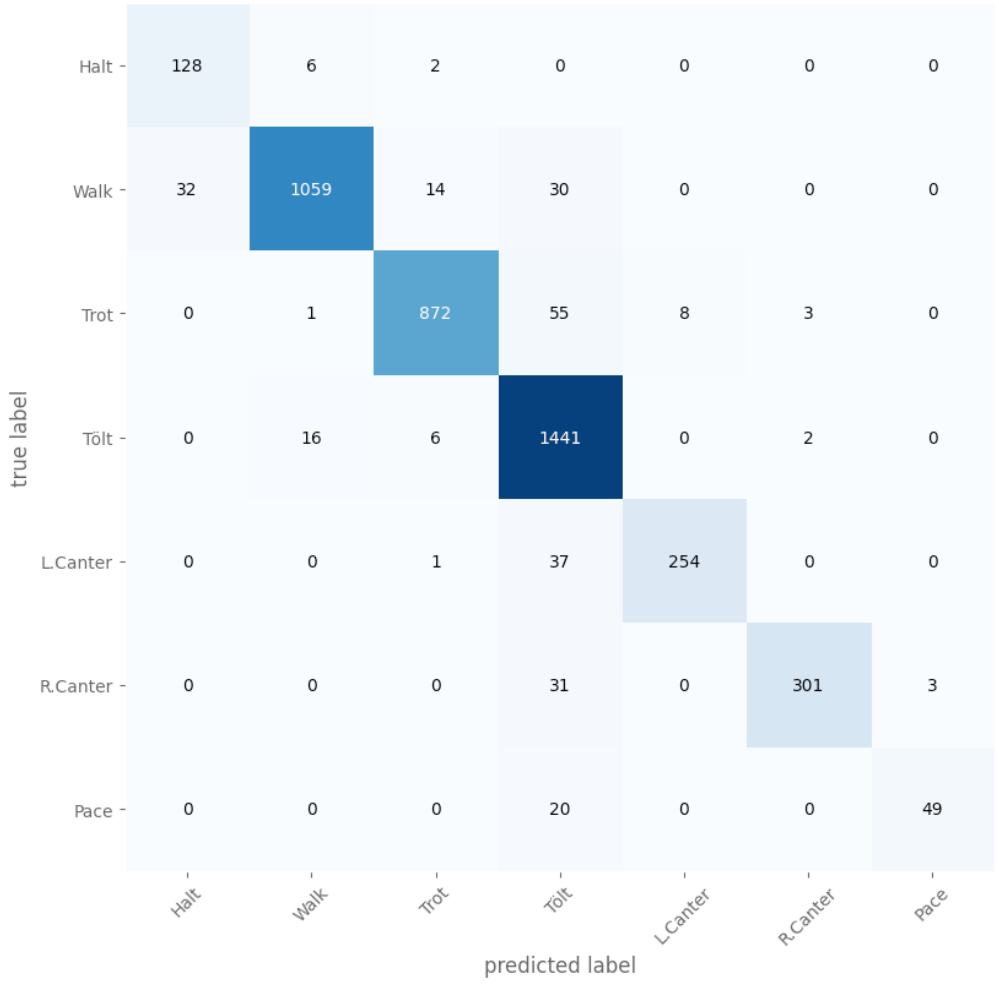

Figure 7: A confusion matrix for the case where both TS and Judge exclusion period was 0ms.

Discussion

We have demonstrated a very high level of agreement between TS and qualified sport judges when classifying the gait of Icelandic horses. The use of exclusion periods around transitions can take the agreement to over 99% but the agreement is around 94% even without any exclusion. We earlier framed this in terms of ‘accuracy’ by considering the judge classifications as the ground truth. In reality we are assessing the agreement between two different measuring approaches, and it is valid to question whether TS is more accurate than human observers in some circumstances. There are often areas where the judges disagree about the gait. For example, a very pacey tölt coming out of canter may be classed as pace by some and tölt by others. There are also cases on the boundary between walk and tölt and

between tölt and trot. It may be interesting to analyse these disagreements as a project in itself.

The main shortcoming of this study is that more instances of flying pace should have been included. If the opportunity arises to study some more pace horses, we will update our findings accordingly and post the results on the TS website[4].

[4] https://toltsense.com

Part 2: Validation of Hoof-on / off Detection

The classification of gait and the analysis of the quality of the beat rely fundamentally on the measurement (or estimation) of hoof-on and hoof-off time. We have already shown that the gait classification is up to 99% accurate when compared with qualified sport judges' classification. But TöltSense goes beyond mere gait classification and offers an analysis of the gait quality in terms of beat and suspension. This section is concerned with quantifying the accuracy of the hoof-on/off event measurement that underpins this feature.

Summary

A five-gaited Icelandic horse wearing a TS sensor kit was ridden and filmed on an oval track. The filming was done using a 100 fps camcorder, and this video was later used to construct a timeline of hoof-on and hoof-off events by hand, which served as the ground truth. Independently, TS logged the whole session and constructed a separate timeline of hoof-on and hoof-off events that was compared to the ground truth. These timelines were temporally aligned with respect to hoof-on events using cross-correlation, and statistical analysis was used to estimate the accuracy of TS. First, the distribution of differences between the video-derived and TS-derived hoof-on timings was found to be approximately normal with a standard deviation of 10.6 ms. Further analysis suggests that the true precision of TS in measuring the timings of hoof on events is 9.4 ± 0.9 ms. A similar procedure was followed to analyse the TS hoof-off measurements. The accuracy of TS hoof-off can be expressed as the mean difference with the corresponding video hoof-off events using the aligned time series. The accuracy was found to be -0.3 ms, and the sd was 10.2 ms.

Gait beat quality analysis relies on measuring the time intervals between hoof-on events.[5] With that in mind, we calculated the timings between successive hoof-on events for both video and TS events for comparison. A Bland Altman analysis showed a mean diff of 0.12 ms with upper and lower limits of agreement of 28 ms.

The results indicate that the beat quality analysis of TS is at least as accurate as an idealised frame-by-frame analysis of a 30 fps video. Although other academic studies in the field achieve much higher accuracy, this is easily sufficient for the real-time analysis of beat and the pursuit of Icelandic horse training for sport.

Materials and methods

For this exercise, a single 5-gaited Icelandic horse was used. The horse was a 16-year-old gelding recognised to have five good gaits and previously represented Great Britain at the World Championships for Icelandic horses. The horse had no history of lameness and was demonstrably sound and healthy immediately prior to data collection. The rider (and owner) of the horse was an experienced amateur who has also represented Great Britain at the same championships.

The horse was filmed riding on a FEIF approved oval track with a firm gravel surface in southern England. The filming was conducted at 100 fps and full HD resolution using a video camera (Panasonic HC-V770). The horse was fitted with a TS sensor set, and the rider was carrying the phone (iPhone 8) in their jacket pocket. The session lasted around 12 minutes, and the filming and TS logging was continuous throughout this period. During this time, the rider was able to demonstrate all 5 Icelandic gaits.

[5] For example, tölt is a symmetric four-beat gait, and the ideal performance requires that each footfall is precisely separated by 25% of the overall cycle time. The repeating footfall pattern is the same as in walk: for instance, HL -> FL -> HR -> FR, etc. If the cycle time were 500ms, for example, then HL would land at 0ms, FL at 125ms, HR at 250ms, and FR at 375ms. The time of 500ms represents two cycles per second, and this happens to be a reasonable cycle frequency for medium to fast tölt. Very fast tölt or flying pace might be performed at 2.5 or even 3 cycles per second, but this is definitely the extreme. Reliable analysis of beat quality requires that the error in the identification of the hoof-on time (and therefore inter-limb timings) is small compared to the cycle time. Therefore this paper aims to quantify the accuracy achieved by TS in identifying hoof-on events and thus the level of confidence we can have in the gait analysis.

The filming began by capturing the act of pressing the “START” button in TS. This action launches the session and marks the start of logging to the app’s database, so filming it offers a quick way to line up the video with the TS log. The video was used for a frame by frame analysis to identify hoof-on and hoof-off times with a custom-built web application. When identifying hoof-on, the first frame where the hoof was clearly in contact with the ground was selected. For hoof-off, the first frame where the hoof was clearly off the ground was selected.

The hoof events were not always clearly visible in the video for a variety of reasons: e.g., boarding at the edge of the track, one leg obscuring another, or the horse being just out of shot. Therefore prior to conducting the frame analysis, periods of clear view were first identified for each gait. Unfortunately, none of the trot shown was usable because insufficient consecutive clear-viewed strides were available. Two viable sections of video for each of the other gaits were identified, and a total of 383 hoof-on events were identified and recorded. The speed of the horse in this data set ranged from approximately 6 to 35 km/h, which corresponds to stride frequencies of between 60 and 140 cycles per minute. The distribution of hoof-on events by gait is shown below.

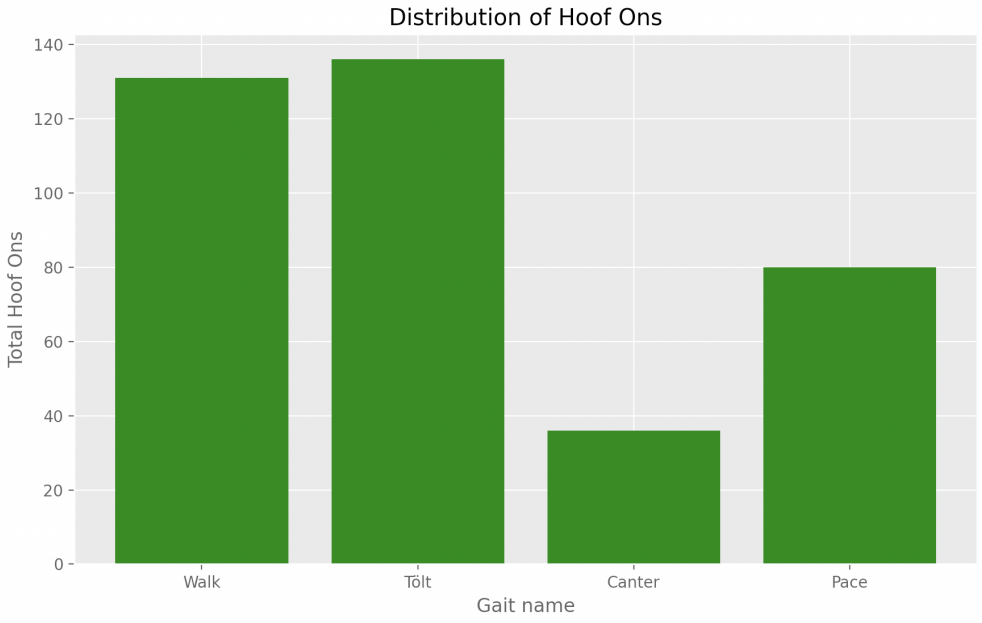

Figure 1: Distribution of hoof-on events by gait.

Data Analysis

The statistical analysis and visualisation of data were carried out using Python 3 and a range of packages, including numpy, matplotlib, pyplot, pandas, json, scipy.stats and statsmodel.api. In the following, I will refer to events logged from the video as ‘video events’

and events logged by TS as ‘TS events’. When referring to the exact timing of events, I use the notation TVid and TTS.

Our study of the rhythm of the footfalls, or the ‘beat’, is primarily concerned with the relative timing of hoof-on events rather than their absolute times. For example, we might be interested in the interval between the front left hoof-on and hind right hoof-on as a fraction of the overall cycle time. Therefore the first step in this method is to align the time series of video events and TS events as closely as possible. The timing of events captured from the video is expressed as the elapsed time in ms starting from zero, whereas the events logged by TS are expressed as a unix timestamp (milliseconds since 1970-01-01 00:00:00.000). First, we subtracted the TS session start time from all subsequent TS event times and then we added on the elapsed time of the video from the point where the START button was pressed (28.92 s). To achieve a more precise line-up, we performed a cross-correlation.

To begin with, both the video and TS series were just lists of timestamps. Each time series was transposed into a new list, the index of each element representing a single millisecond increment and given a value of zero. For each TS event, a value of 1 was applied at the index corresponding to the time (e.g., a time of 1000ms would be a value of 1 at index 1000). The video events were treated similarly except that instead of a single 1 value at the time index, an isosceles triangle was plotted with the apex at the time set to 1 and a base width of 30 ms (see Figure 2). The cross-correlation was then performed by sliding the TS series across the video series from -2000 to +2000 ms and calculating the sum of the element-wise product of the two series. The offset at which the greatest value was produced was then used to set the ideal offset for the analysis; this was found to be 38 ms. A visual inspection confirmed that this was an excellent fit (see Figure 3).

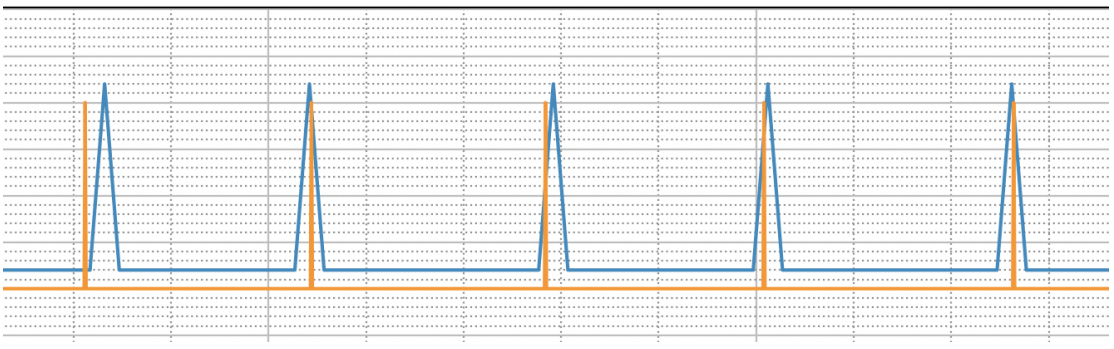

Figure 2: Here is an example of the aligned TS and video plots. The TS events are orange and the video derived events are in blue. The purpose of the triangle is so that partial alignments also contribute to the overall value.

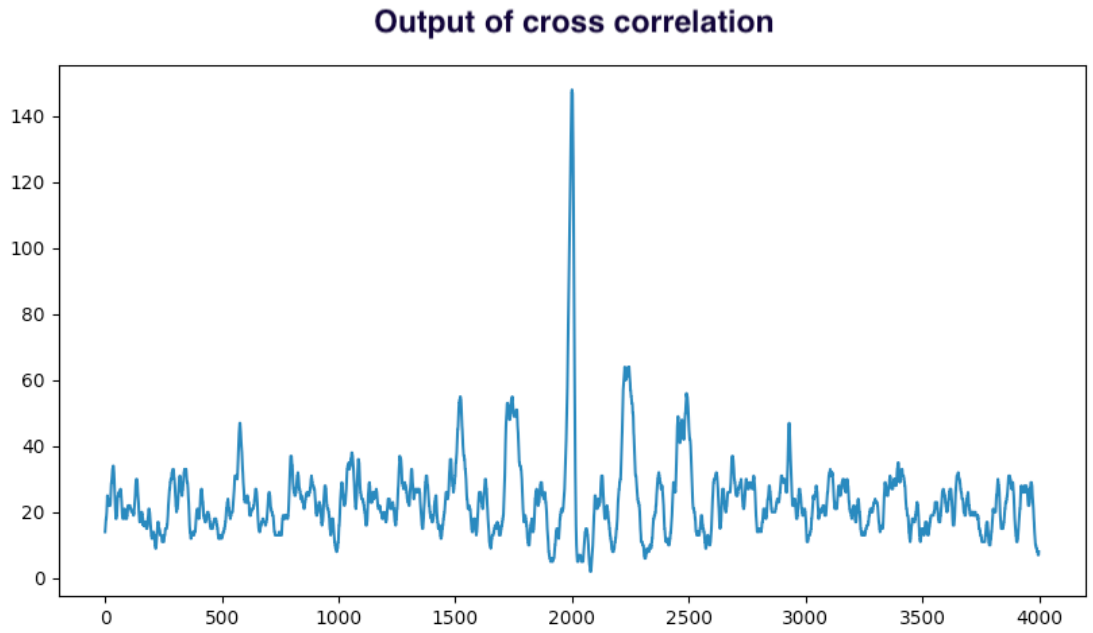

Figure 3: A representation of the output of the cross correlation. The strong sharp central peak shows that the alignment is correct. Note that the x-axis corresponds to shifts from -2000 ms to +2000 ms.

Once the optimum alignment was achieved, it became possible to investigate how closely the TS events lined up with the video events. The difference between the respective timings was calculated throughout the entire range in preparation for analysis. This preparation allows us the possibility of estimating the precision of the individual event timings

Since we are interested in comparing the two methods of measuring inter-limb hoof-on timings, we also calculated the time difference between successive hoof-on events for both series. Both series were sorted by timestamp and then for each event the time interval between it and the one before was found. Since the data set consists of short sections of focus interspersed with longer periods without analysis, some lengthy invalid periods arose, which were then discarded. These pairs of measurements were used for a Bland-Altman analysis to estimate the limits of agreement between the 100fps video analysis and the TS automated event detection with respect to inter-limb hoof-on timing.

Results

A total of 383 hoof-on events were identified from the video and paired up with their corresponding events from the TS log. For each pair, a difference (T_{Vid} - T_{TS}) was calculated, and the mean (-2.6 ms), mode (4.0 ms), and standard deviation (10.6 ms) were derived (see Figure 4).

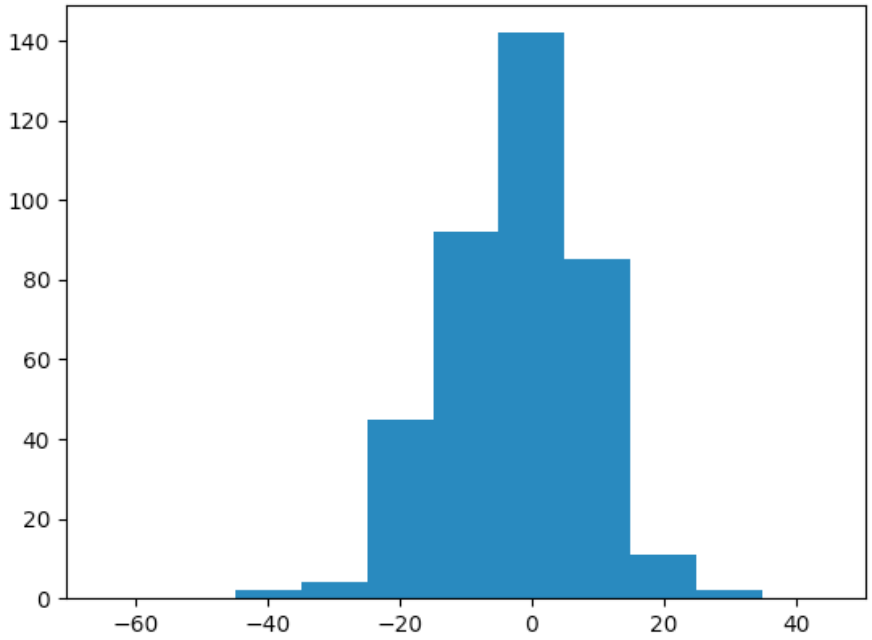

Figure 4: A histogram of the differences between TS events and Video events.

A similar exercise was carried out for each gait and leg of the horse. No significant gait, leg, or speed dependence was observed for hoof-on timings.

TS is concerned with measuring the beat quality by comparing inter-limb timings with overall cycle periods. The two lists of successive hoof-on intervals were therefore used to generate sequences of inter-limb hoof-on timings. These were used to generate a Bland-Altman plot, which allows us to assess how well the two systems of inter-limb timing measurement agree (see Figure 5).

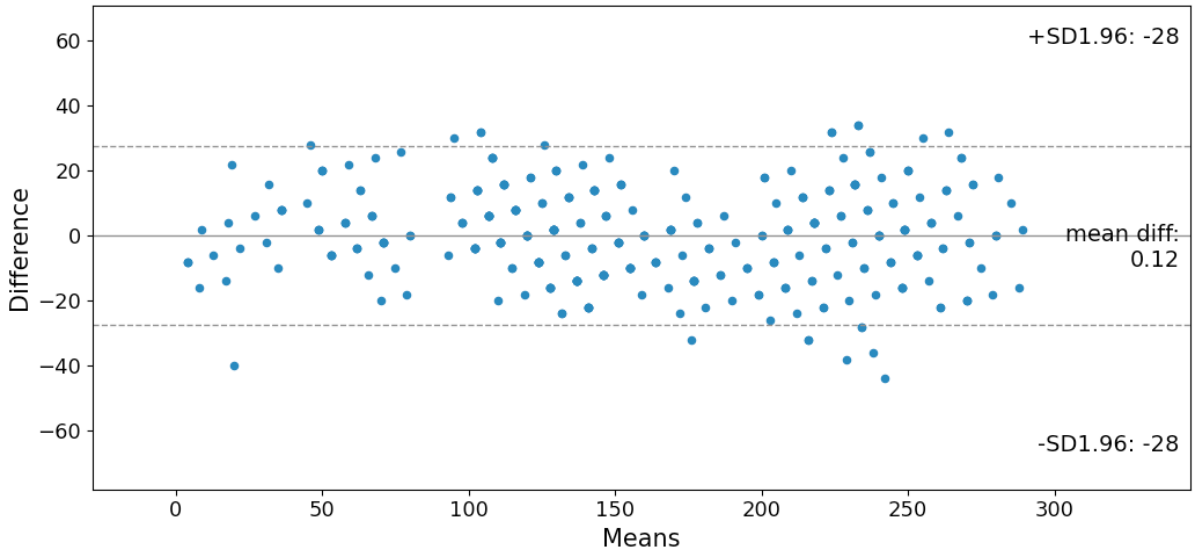

Figure 5: Bland-Altman plot comparing two methods of measuring inter-limb timing

The mean difference was calculated to be 0.12 ms and the upper and lower limits of agreement were found to be +28 ms and -28 ms respectively.

Discussion

The histogram in Figure 4 represents the distribution of differences between our two measurement methods with regard to hoof-on events, but what we really want to know is how the TS measurements are distributed around the true hoof-on times. Both video analysis and TS have a degree of uncertainty in their measurement of hoof-on times, and the distribution of differences in the histogram reflects the combined uncertainties of both methods. Suppose we can quantify the uncertainty in the video-based measurements of hoof-on and combine this with a statistical analysis of the diffs T_{Vid} - T_{TS}. In that case, we should be able to estimate the uncertainty in the TS hoof-on measurements.

The approach is as follows:

1. Assume the premise that the distribution of diffs is solely the result of the combined uncertainties in the video analysis and the TS hoof-on measurement methods.
2. Derive best and worst-case uncertainty values for the video-derived hoof-on times.
3. Analyse the collection of diffs to work out upper and lower estimates for the standard uncertainty in that distribution.

4. Assume that the TS hoof-on measurements are normally distributed around the true hoof-on events.
5. Work out using computational modelling what standard deviation in the TS measurements would be required to yield the observed distribution of diffs.

We can estimate the uncertainty in the video-derived timings with a model based on the uniform distribution. In the ideal scenario, when performing video analysis to identify hoof-on, we can see one frame where the hoof is not yet touching the ground, and then the next frame shows that hoof-on has clearly occurred. If the first frame time is *a* and the second *b* then we can say with certainty that the true hoof-on time occurred between bounds *a* and *b*, but we cannot say where it is most likely to have occurred. That is, when analysing a data set after the fact, we have to admit that all times between *a* and *b* are equally likely to have been the 'true' hoof-on. For such a uniform distribution, we can say that the best estimate of the hoof-on time is (*a* + *b)* / 2, and the standard uncertainty is given by

$$\frac{b-a}{\sqrt{12}}$$

In our case *b* - *a* is 10 ms, so the uncertainty comes out as 2.9 ms. Note that this is the best case uncertainty, where for every frame time selected, the true hoof-on falls within the 10 ms leading up to it. This is almost certainly not the case in practice, however, and there are likely a fair proportion of identifications that were one frame too early or one too late.[6]

In order to construct a reasonable worst-case uncertainty estimate, let us assume that 10% of the selected frames were one frame too late and 10% were one frame too early. This means that for each value recorded, the likelihood of the true time falling in the 10 ms leading up to that time is 80% and the likelihood of the true time being between 20 and 10 ms earlier is 10%, and the likelihood of it being in the 10ms following it is also 10%. We can then model the distribution of true hoof-on relative to the measurement as three adjacent uniform distributions with different percentage likelihoods - 10:80:10

[6] It is also worth noting that we have not addressed the question of what really constitutes the 'true hoof-on'. When a hoof lands, it is not a single instantaneous event; rather, it is a process. It might start when the first part of the hoof or shoe makes the slightest contact with the ground, and later the whole of the hoof is in contact with the ground, and the leg is fully bearing weight.

We estimate the standard error by sampling repeatedly from a distribution where a value in the set {-19,-18,…,-10} ∪ {1,2,…,10} is selected uniformly at random with a 20% probability and a value in the set {-9,-8,…,0} is selected uniformly at random with an 80% probability. With a 100,000 samples the standard deviation was found to be 5.3 ms. This result allows us to estimate the error in the video event times to be between 2.9 and 5.3 ms.

Turning to the distribution of diffs mentioned above, we determined through experimentation that the sd, which is generally used as the standard uncertainty for normally distributed data, was 10.6 ms. However, we can see that the distribution is not perfectly 'normal' and is negatively skewed with a mean diff of -2.6 ms and a mode of 4 ms. We must ask what is causing this skew. If we could directly compare the TS hoof-on times with the true values, we would expect to see a normal distribution. After all, there is no upper or lower bound for the timestamp that the algorithm could return, and the sensors are capturing kinematic data resulting from the actual event we are trying to detect. Since this skewed distribution results from the combined errors of both measurements, it is likely that while the TS measurements are indeed normally distributed around the true value (or an offset from it), the video measurements are likely not fitting the idealised uniform distribution described above. Possibly this is because where the optimum frame is not selected, an observer is more likely to select a late frame than an early one because it is easier to detect when a hoof is far from the ground than when it is close to it. This might explain the skewed distribution.

Another way to estimate the standard error in the distribution of diffs is to calculate the proportion of diffs within a given range around zero. We would expect approximately 68% of the measurements to be within 1 sd of the true value for a perfectly normal distribution. In this dataset, it turns out that the percentage of diffs within -10 to +10 is 70%. Since more than 68% of the values lie within this range, we might reasonably say that the true uncertainty is at most 10 ms.

We can summarise the best and worst estimates for the uncertainty in T_{TS} and T_{Vid} - T_{TS} in the following table.

Measure	**Best estimate**	**Worst estimate**
T_{Vid}	2.9 ms	5.3 ms
T_{Vid} - T_{TS}	10.0 ms	10.6 ms

The distribution of the differences (T_{Vid} - T_{TS}) results from the uncertainties inherent to both measurement methods. Armed with the estimates of the upper and lower bounds of the uncertainty in each, let us try to estimate the upper and lower bounds of the uncertainty of T_{TS} relative to the true hoof-on.

Again, let us take an empirical computational approach. To create a simulated collection of diffs to analyse, let us imagine we are repeatedly measuring a true hoof-on event with time T using our two methods which yield T_{Vid} and T_{TS}.

As before we can define a list *s* that holds sample differences. For a large number of iterations we sample from whichever distribution we are using to model T_{Vid}. This value ΔT_{Vid} represents the difference between the true hoof-on and the value measured by the video, T - T_{Vid}. Next, we select a random number from a normal distribution defined by mean = 0 and a target sd, *sdt*, which is a potential sd value for the TS measurements. This number ΔT_{TS} represents the difference between the true hoof-on and the value measured by TS, T - T_{TS}. We then sum these deltas and add the result to the list *s*. Finally, after many iterations, we can calculate the sd of the resultant list of diffs. Repeating this process while varying the standard error estimate for T_{Vid} - T_{TS} and also for the distribution of T_{Vid} allows us to discover what standard error in the TS measurement would result in the distribution of diffs that we discovered before. In other words, given an estimated uncertainty in T_{Vid} and measured standard error in T_{Vid} - T_{TS}, we can use maximum likelihood estimation to find what value of *sdt* would be required to generate a distribution that matches.

First, let's use the best-case single uniform distribution for modelling T_{Vid} with sd = 2.9 ms and also that the sd in the diffs is 10.6 ms. The question is, what standard error in T_{TS} would be required to achieve a sd of 10.6 in the diffs? The answer, found by trial and error, is 10.2 ms. On the other hand, suppose we accept the worst-case 10:80:10 composite uniform distribution for the T_{Vid} values and also that the standard error in the diffs is 10 ms. In order to achieve a distribution with sd = 10 ms, a *sdt* value of 8.5 ms is required.

Taking an average of the upper and lower estimates for the sd in T_{TS} we can say that the true uncertainty is 9.4 ±0.9 ms. To put this into context, we can make comparisons with frame-by-frame video analysis. As discussed earlier, performing an idealised video analysis gives you timestamps which mark the upper bound of a uniform distribution containing the

true value. If δt represents the frame interval, then the standard error in such a measurement is given by δt / $\sqrt{12}$. For example, with 100 fps video, the interval δt is 10 ms, so the standard error is 2.9 ms. We can therefore calculate the fps values that correspond with the upper and lower estimates for the sd of Tts, and we find that the values are 28.3 fps and 34.0 fps. This enables us to say that the fps equivalent is approximately 31 ±3 fps.

To summarise, we produced reasonable estimates for the best- and worst-case uncertainties in the measurement of hoof-on using 100 fps video and the distribution of diffs between the video and the measurements of TS. Using the assumption that the TS measurements are normally distributed around the true hoof-on events, we have shown that the uncertainty in any given measurement using TS is 9.4 ±0.9 ms. To put this into context, that is roughly equivalent to the accuracy you would get from performing an idealised frame by frame analysis of 30 fps video.

Analysis of hoof-off accuracy and precision

Beat quality is a function of the hoof-on timings. However, other important factors relating to performance and gait quality can only be determined by measurement of hoof-off: for example, duty factor and suspension. The duty factor is the proportion of the cycle that a given leg is bearing weight. Strictly speaking, we are not able to measure whether (or how much) weight is borne by a leg using video or leg-mounted sensors: we can only infer whether a hoof has made contact with or has been raised off the ground. Nevertheless, dividing a stride into a stance phase and a swing phase delineated by hoof-on and hoof-off times is a pretty good way to arrive at an estimate of duty factor. Full suspension is defined as the period in the gait cycle where all legs are in the swing phase, and none are in contact with the ground. There is no suspension in a walk, but this measure is relevant to trot, flying pace, and canter. In tölt (when the beat is clean), there is sometimes a small amount of full suspension but only at very high speeds. However, we are very interested in so-called ‘half suspension’, or ‘front-suspension’: a horse showing good tölt should be light in the forequarters and exhibit two phases in each cycle where both front legs are off the ground. This will normally also be reflected in the duty factor measurements because the forelegs are spending less time on the ground than the hinds.

As with the hoof-on timings, the log of hoof-offs from TS was compared with the hoof-offs recorded from the video. In other words, for each video event, the diff with the TS event was

calculated and added to a list for analysis. The video was already optimally aligned with the hoof-on times from TS and so making this comparison is directly assessing the accuracy of the TS hoof-off times relative to the hoof-on times.

A total of 347 hoof-off events were included. Overall the mean difference between TS hoof-off and the video log was -0.3 ms, and the diffs had an sd of 10.2 ms. The mean difference of -0.3 ms shows that the accuracy of the hoof off measurement is excellent. The standard deviation is slightly lower than that found in the hoof-on analysis, and it was also found that 69% of diffs were within 9ms of zero.

Conclusion

This validation study was in two parts. First, we examined the accuracy of TS gait classification by comparing TS's log with the opinions of four qualified Icelandic sport judges. The data set included eight different Icelandic horses and all 5 Icelandic gaits.We showed that TS is over 99% accurate at gait classification by allowing a two-second exclusion period around transitions. Without the exclusion period, the result is still around 94%. Second, we explored the accuracy of TS's hoof-on and hoof-off measurements by comparing the TS log with timings derived from 100 fps video analysis. The aim was to quantify the accuracy and precision of the interlimb hoof on timings and stance durations measured by TS. We drew a comparison with an easily understood alternative method of establishing these metrics: namely, video analysis. We showed that TS is at least as accurate as the best possible outcome that could be reached with a frame-by-frame analysis of 30 fps video.

Video analysis is routinely used in both teaching and assessment. For example, a trainer will commonly use a video to demonstrate or communicate a certain point to their student. It is also now common to find 'online' competitions where riders submit videos of themselves doing a test with their horses. Judges have no problem assessing a show that was filmed at 25 or 30 fps, and by definition, they can rate the quality of the beat in all gaits from such video. Also, when judges are being trained, it is customary to explain concepts like Lateral Advanced Placement using video analysis. On this basis, I think it is fair to claim that TS is as good as any human observer when it comes to classifying the gait and analysing the

beat.[7] However, I would argue that human observers are probably not very effective when it comes to quantifying metrics like suspension duration or duty factor, so this is one area where TS is probably superior to human observers.

This study could be improved and extended in several ways. First, more horses could be used, particularly for the frame-by-frame video analysis. Second, more flying pace should be included in the gait classification study. Third, a higher frame rate could be used for the video frame analysis. Fourth, it might be useful to assess the accuracy of TS on different surfaces. Fifth, the lack of trot in the video analysis was unfortunate, and a repeated study should definitely include this.

[7] I do, of course, accept there are many other aspects of a performance that TS cannot capture or evaluate