MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 3 - 4, 2017 HYNES CONVENTION CENTER

Mining Muscle Use Data for Fatigue Reduction in IndyCar

Yasuyuki Kataoka, Douglas Junkins
NTT Innovation Institute, Inc.
Email: kataoka.yasuyuki@ntti3.com, Doug.Junkins@ntti3.com

Abstract

This paper discusses how data analytics on muscle use in extreme racing conditions can be conducted to find actionable insights for the driver. One of the important insights is how to minimize driver's muscle fatigue during a race, because IndyCar has regulations forbidding the use of power steering. This paper tackles two technological challenges: 1. data validation on noisy signal obtained from wearable device in extreme condition, 2. data cultivation to find actionable insights for the driver from heterogeneous racing data. First, we propose a data quality assessment technique, enabling the judgment of whether data is reliable or not. This qualitative analysis revealed that the data validation method works 99.5% accuracy to classify data as reliable or not. Second, we propose a data visualization tool based on unsupervised learning that enables the driver or mechanics to discover useful feedback. We identified and demonstrated several actionable insights, e.g., identifying potential relaxation points.

1. Introduction

IndyCar is an American-based auto racing sanctioning body for Championship auto racing. Unlike other racing formats, such as the Formula One, IndyCar has regulations forbidding the use of power steering. This requires drivers to exert more force on their forearms, which dramatically deteriorates their performance as their muscles fatigue during a race. Hence, saving a driver's muscle use during a race is a beneficial insight for the driver. This paper tackles this challenge: how data analytics can improve driving performance, that is, minimize the use of forearm muscles during a race.

In the research of auto-racing, various approaches are conventionally taken to improve driver's performance or safety. One approach is the trajectory path optimization based on the driver's record.[1] The findings from trajectory analysis are used for the path planning of self-driving cars.[2] Another approach is a real-time decision system for tire changes within a race.[3] Moreover, there is a research for driver's safety. One approach is around heat prevention using a temperature sensor on the driver.[4] However, in our survey on publicly available papers, no research focused on analyzing forearm use during race with the consideration of heterogeneous data.

To improve the driver's performance with the focus on muscle use, this paper tackles two major technological challenges:

A. data validation on noisy signal obtained from wearable sensors in extreme condition,
B. data cultivation to find actionable insights for the driver from heterogeneous racing data.

For challenge A, one of the common challenges for wearable devices is data quality and validation. The quality of the signal coming from wearable device is very sensitive to whether or not it is properly attached to the body. Thus, for challenge A, we propose a data quality validation technique that enables the judgment of whether the data is reliable or not. This methodology is based on the comparison between actual Electromyogram (EMG) and predicted EMG, which is computed by a Machine Learning technique. If the actual EMG deviates significantly from the predicted EMG, the actual EMG is considered not to be valid. To guarantee the performance of the prediction, the feature in the prediction model uses only the car's telemetry information, i.e, excluding the EMG information itself as a feature. This is because the EMG signal itself may be too noisy to use for prediction, while the car's telemetry information can provide stable signals. Our calculations show that this qualitative analysis based data validation method works with 99.5% accuracy to classify the data reliability.

For challenge B, a data visualization and interaction tool was designed that enables the race team to cultivate heterogeneous data and discover useful insights in an intuitive manner. The computation method behind this tool is a multi-modal analysis of EMG and car telemetry data. This analysis is unsupervised learning using the following technique; 1. cluster data points in a geographical fashion, 2. find similarity between EMG and the car's telemetry data. Based on this analysis, locations are identified where the driver exerts unnecessary force during the race, in other words, where the driver may be able to rest and recover.

The work in this paper is based on data collected over several races during the 2016 Verizon IndyCar Series from Tony Kanaan, a driver from the Chip Ganassi racing team. To measure the driver's muscle use, we used a special wearable fabric called hitoe[1] [5], to collect the driver's EMG data from their forearms during the race. To meet the IndyCar requirement for flame retardant fabric, these hitoe sensor patches were sewn inside the driver's standard Nomex undershirt. Along with the EMG data, the car's telemetry information was also collected, including three-axis accelerometer data, steering, brake pressure, gps location data, complete with time stamp and lap numbers. This heterogeneous dataset is comprehensively utilized. We believe this work is the world's first publication to collect and analyze forearm use during real-world racing conditions.

This paper is composed as follows. In section 2, we describe how we collected data from both the car's telemetry and the driver's vital information. In section 3, the data quality validation challenge is discussed including the methodology and the evaluation. In section 4, the data cultivation and visualization challenge to identify actionable insights is discussed with both the methodology and evaluation. Finally, in section 5, we conclude this paper.

2. Data Collection

In this section, the collection and structure of the data used in this analysis is described.

2.1. Data Collection

A variety of data points are collected while the driver is on the race track. The hitoe fabric is sewn inside the driver's long sleeve Nomex undershirt and has two sets of sensors collecting bioelectrical signals. One set is used to collect Electrocardiogram (ECG) with sensors located around the rib cage and the second to collect signal for EMG with sensors located around the forearm. The ECG data was

[1] hitoe is a wearable fabric developed by NTT Group. The ordinal product of hitoe can measure ECG through API. In addition to it, we added EMG measurement functionality for this work.

not used for the analysis described in this paper. The sensors communicate through a bluetooth receiver connected to the onboard telemetry system. The receiver captures data at 200 samples per second. Each race offers different number of laps ranging between 50 to 300 laps with an average distance of 2.2 miles per lap.

In addition to the driver's wearable sensor information, we collect from the onboard telemetry system accelerometer data (latitude, longitude, vertical), steering angle, speed (mph), throttle pressure, brake pressure, engine rpm, angular acceleration, GPS coordinates (relative to a fixed point on the track), and seconds of gap between the car ahead and behind the driver. This information is collected through the onboard sensors provided by the Chip Ganassi Racing team. Each data point is collected at a different sampling rate depending on its need for the race strategy.

Data is collected through a private network that is accessible during the race allowing the team to do near-real time analysis. However, due to limited bandwidth, the transmission is limited to lower sampling rates, which didn't provide the required granularity for the analysis. This paper focuses on post-race analysis which allows higher frequency data to be used, which is downloaded from the onboard telemetry system after the race.

Along with all data points, we collect timestamps from the onboard real time clock (RTC). The RTC is used to anchor data collected across differing frequencies. Using this RTC channel value, data from all channels is realigned to the highest frequency channel, while listing a blank value where the source channel collected data at a lower frequency. This is the base data preparation is conducted before subsequent analysis as documented in Section 3.

2.2. Dataset for wearable sensor's quality analysis

One of the challenges that this paper tackles is the validation of sensor data, or, in other words how can we judge the cleanliness of the data? This argument involves two different questions.

The first question is whether the wearable sensor, hitoe, has capability to capture EMG during extreme racing condition. This may require further analysis by physiological experts to guarantee that our wearable sensor has the capability to capture the signal. This is out of scope of this paper's focus since we assume the sensor has the capability to collect EMG data based on the following observations. A sample of EMG data is shown on the track with time stamp in Figure 1. The left bottom corner is expanded in Figure 2 where the radius of the circle represents the amplitude of EMG signal. As Figure 2 shows, it makes sense that high EMG readings occur just before the corner after a long straight section of track. Therefore, in this paper, the wearable sensor, hitoe, is assumed to function properly even in extreme racing condition.

The other question is whether the wearable sensor attaches to the driver's body reliably. Even if the sensor has the capability of measuring EMG, it cannot provide a clean signal if it is not attached to body properly. This is the data validation problem that this paper focuses on. Namely, our system addresses the attachment question by classifying the sensor data as clean or dirty.

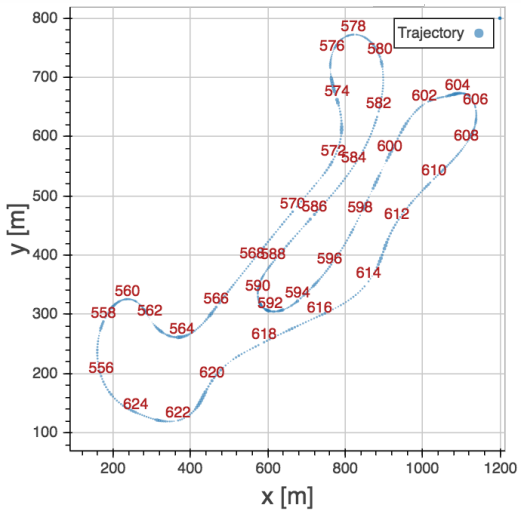

Figure 1. EMG behavior on entire track

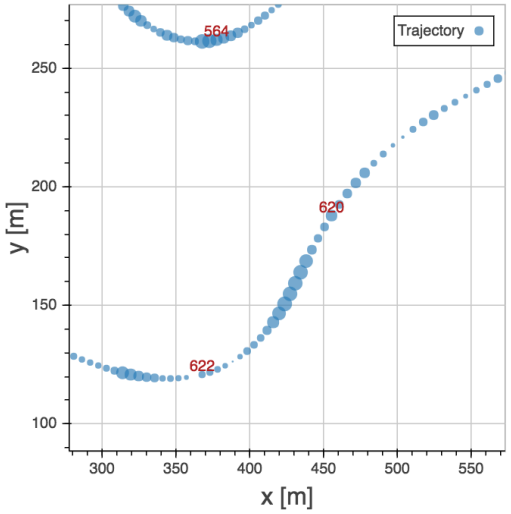

Figure 2. EMG behavior on corner after long straight

To evaluate the classification performance of our system, the labeled data or ground truth, needs to be determined. To do so, we collected data in two different ways with the hitoe wearable fabric. One piece of fabric is worn with tight compression for clean data. While tight compression collects clean data, the downside of it is discomfort to the driver. Therefore, it not realistic to use tight compression in real competition. The other fabric is worn with loose compression for less reliable data. While loose compression is more comfortable for the driver, it produces noisy data during extreme racing condition. We collected these two different types of data over same course. The details of these datasets are described in section 3.2.1. The challenge for reliable data collection is finding the balance between compression that is tight enough to provide clean data, but loose enough to be comfortable during a long race. This balance point will be dependent on the driver.

3. EMG Data Quality Analytics

One of the common challenges for wearable devices is data quality and validation. The quality of the signal coming from wearable devices are very sensitive to whether it is properly attached to the body or not. Especially in the case of IndyCar, the EMG data is highly affected by the extreme forces experienced in racing conditions. If the data is not valid and reliable, it may lead to faulty analysis and incorrect insights. Thus, it is important to validate the data quality before proceeding to advanced analysis.

3.1. Methodology – Machine Learning on Heterogeneous Data

The methodology for data validation is based on the comparison of actual EMG data and predicted EMG. If the driver's actual EMG significantly deviates from the predicted EMG, the collected EMG is not considered valid. The key to this methodology is how to predict the EMG value with high accuracy in a reliable manner via Machine Learning. Our hypothesis is that we can achieve this by using only heterogeneous and reliable car telemetry data as features for the prediction model, i.e, excluding the collected EMG information as a feature. Our evaluation reveals that this prediction using heterogeneous data has acceptable accuracy. We conduct this data validation process for each lap of the race course since further analysis depends on lap data.

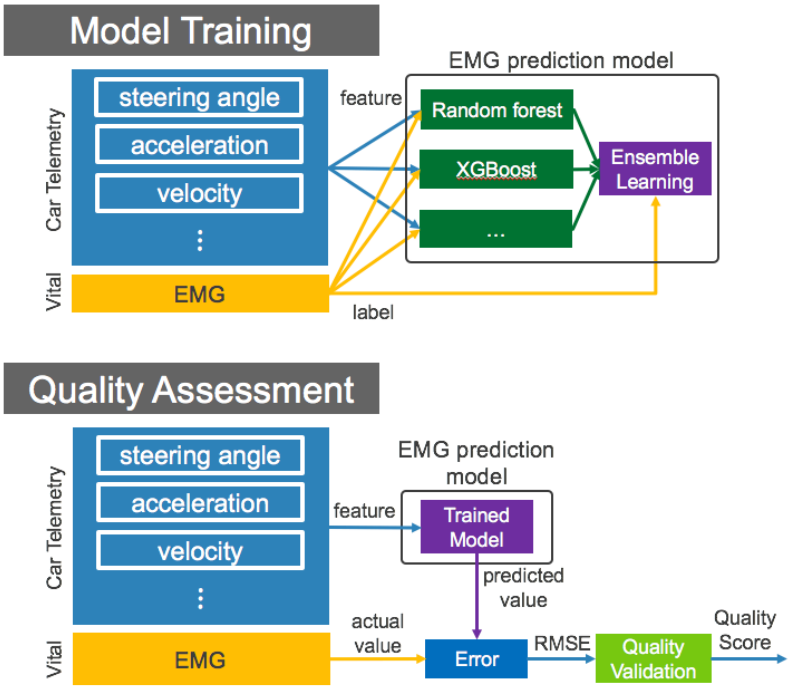

Figure 3. Data Validation Process by Ensemble Learning on heterogeneous time-series data

3.1.1. Step 1: Training Model

In the first process, our EMG prediction model is trained via machine learning. Ensemble learning aggregates different prediction algorithms such as Random Forest, XGBoost,. This way, our model can leverage diverse prediction models to produce more accurate results.

The features are simply designed by car telemetry data. The EMG data is not used as a feature since it is highly affected by how the wearable device is attached to body. In the extreme conditions of IndyCar, temporary detachment from the driver's body can easily happen. Thus, to create the EMG prediction model, we only leverage reliable and heterogeneous data.

This training process requires the clean labeled EMG data which is described in section 2.2. We use only datasets obtained from firmly attached wearable fabric for this training phase.

3.1.2. Step 2: Comparison of predicted value to actual value

In the second process, the quality of the EMG data is measured and judged whether it is usable for further analysis. The data quality assessment is based on the error between actual EMG values and predicted EMG values. We are interested in validating data quality lap by lap for the next level of analysis, which is articulated in section 4.

First, predicted EMG is computed by using the model trained in step 1. Again, the features for this model are only composed of car telemetry information which provide stable signals. Thus, these predicted EMG values are assumed to be able to be acceptably accurate.

Second, the error between predicted value and actual value is computed as

$$e_l = \frac{\sum_{k=0}^{N_l} \sqrt{\left(y_k^{\{p\}} - y_k^{\{a\}}\right)^2}}{N_l} \quad (1)$$

where l represents the lap index, e_l represents the average error (RMSE : Root Mean Squared Error) in lap l, N_l represents the number of data point of EMG in lap l, $y_k^{\{p\}}$ represents the predicted EMG at data index = k, and $y_k^{\{a\}}$ represents the actual EMG at index k.

Lastly, the quality validation will judge if the data at lap l can be used for advanced analytics or not. This classification is based on threshold against e_l. This threshold is experimentally setup.

3.2. Evaluation – EMG Prediction and quality assessment performance

This section shows the two evaluations: 1. The model evaluation for EMG prediction using car telemetry data, and 2. The quality validation by classifying data as dirty or clean.

3.2.1. Evaluation 1 : EMG Prediction Quality

The objective of this evaluation is to investigate the performance of the EMG prediction. The dataset for this is the clean dataset described in section 2.2. This dataset has 40 laps in 10 different practice runs which are composed of 1,655,102 data points. The evaluation uses the 5-fold cross validation method as shown in Figure 4. Each segment is divided by laps. Also, the prediction model is constructed using 10 different sizes of dataset from 10% to 100% for analytics experiment.

The results of the EMG prediction are shown in Figure 5. The best prediction error against the test dataset is about 0.220, while the error against the training data is 0.088 with 100% of the data. Note that the EMG data is scaled by standardization from the original data, meaning the mean equals 0 and the standard deviation equals 1.0. Therefore, this model results in approximately 22% error in the predicted EMG signal. This prediction outperforms a random prediction model.

This prediction performance can be improved by further refinement of the machine learning model. Also, as Figure 5 shows, the error gets smaller when the size of dataset increases. Therefore, the model can be improved by using a larger dataset.

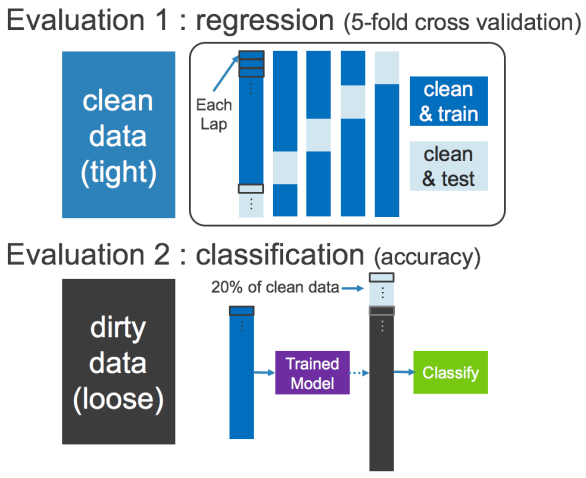

Figure 4. Datasets used for evaluations

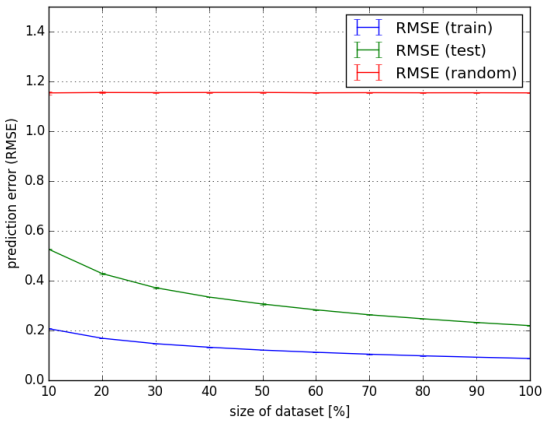

Figure 5. EMG prediction performance (RMSE)

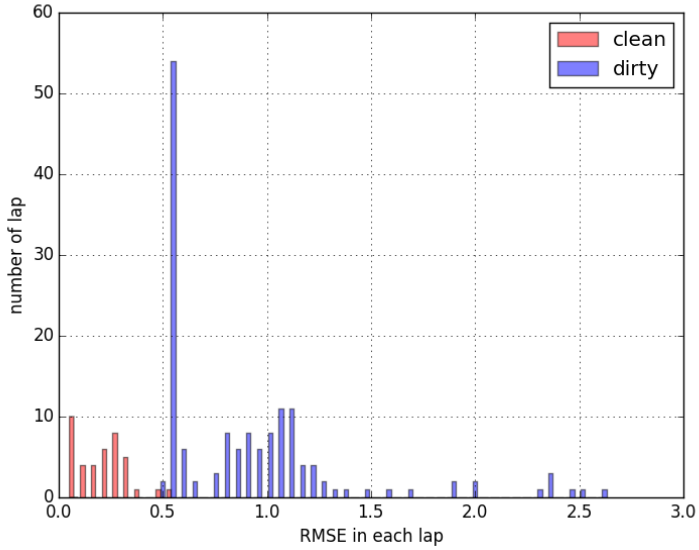

Figure 6. Histogram of lap-based RMSE for both clean and dirty data

3.2.2. Evaluation 2 : clean / dirty classification

The objective of this evaluation is to investigate the performance of the classification of whether data is dirty or clean. The dirty data has 151 laps in 13 different runs, which are composed of 3,574,847 data points. As Figure 4 shows, one evaluation targets 20% clean test data and the second evaluates 20% of the dirty data. This process is repeated five times. This evaluation accurately judges the data quality validation performance.

The result of data quality validation is shown in Figure 6. This figure shows the histogram of RMSE for both clean and dirty classification. Note that this figure shows all results obtained by 5 repeated evaluation at once. The RMSE of the clean data is relatively small because the EMG prediction performs well for clean data. However, RMSE of dirty data varies widely and is relatively large since the predicted EMG value deviates from the actual EMG value. Because the car telemetry data is not likely to be affected by noise, the prediction should roughly perform within 22% error. Therefore, the actual EMG value is considered to have the abnormality.

If the classification threshold is set to be the minimum value of the RMSE of dirty data, 0.507 in this case, the accuracy of classification becomes 99.48%. Realistically, the threshold should be set up conservatively. Although this causes some loss from the clean dataset available for further analysis, including noisy data for further analytics could lead to faulty analysis, which is far worse than losing some valid data.

4. EMG Actionable Insight Analytics

One of the common challenges of wearable devices is discovering actionable insights rather than merely monitoring vital data. In this section, we describe our methodology that identifies the potential points where driver can improve performance. This methodology is focused on using unsupervised learning on identifying the EMG correlations with various data points from the car's telemetry information along various GPS coordinates along the race track. We believe that the

subsequent analytics and visualization should be easily manipulated by the race team so that they, as domain experts, can cultivate the data themselves. Thus, we have provided an interactive data visualization tool that has proven itself valuable in identifying actionable insights.

4.1. Methodology – clustering based analysis

The overall data processing flow is described in Figure 7. First, a signal filter is applied to the EMG to prevent the unreasonable or spiky noise. Second, *k*-means clustering is applied to GPS location to aggregate the data from multiple laps in the same vicinity. Third, the similarity between normalized EMG and car telemetry data is computed at each clustered location. Finally, the analysis result is visualized on an interactive tool.

4.1.1. Filter Design

The objective of the filter is to remove unreasonable or spiky noise. As shown in figure 8, the original EMG signal has spiky noise within 0.1 seconds with the interval of about 0.7 seconds. This appears to be due to the sensor measuring the drivers pulse within the EMG data. For this paper, automatic nervous system such as vasomotor nerve should be ignored. Thus, the spiky noise occurring within 0.1 seconds should be removed. Adjusting parameters, we chose Chebyshev type2 for this case, because it performs well as shown in Figure 8 among other candidates. Other candidates are Butterworth, Chebyshev type1, Elliptic, Bessel, FIR (hamming window). Although Elliptic filter seems to be a little bit better than the Chebyshev type2 to reduce the amplitude, it also causes phase difference. Thus, we chose Chebyshev type 2 for this study.

Note that the chosen filter may not be perfect. Choosing better parameters or using an adaptive filter may improve the performance, however, since this paper's focus is not the filter, we accepted the performance of Chebyshev type 2.

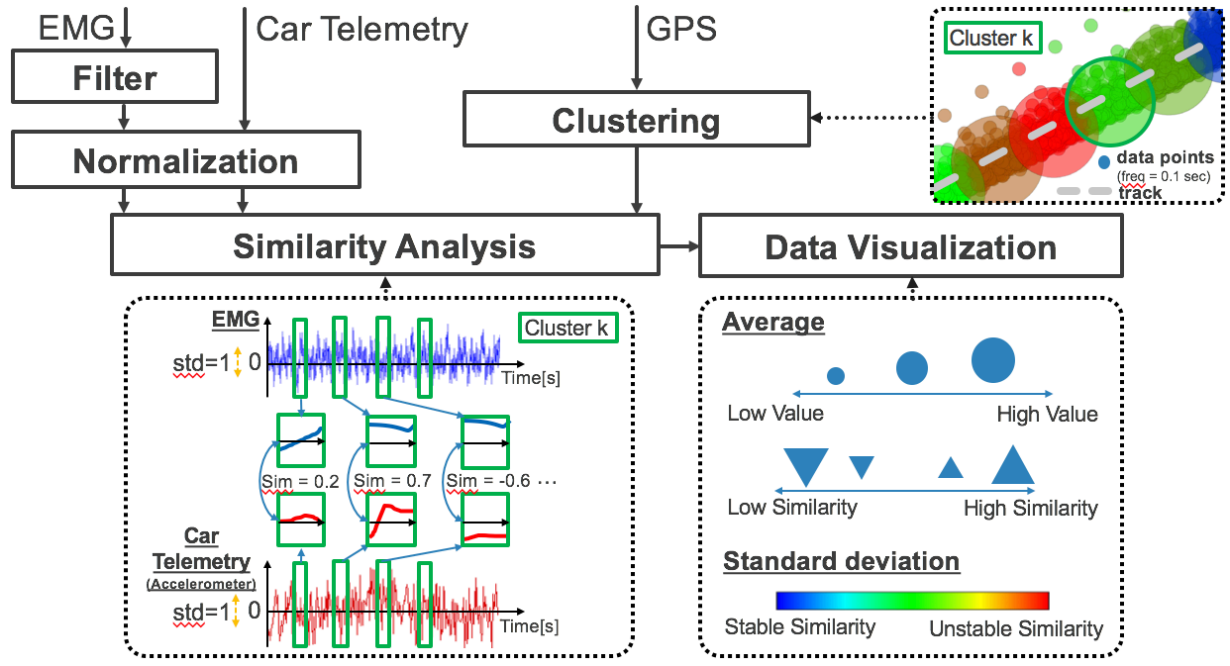

Figure 7. data pre-processing and clustering

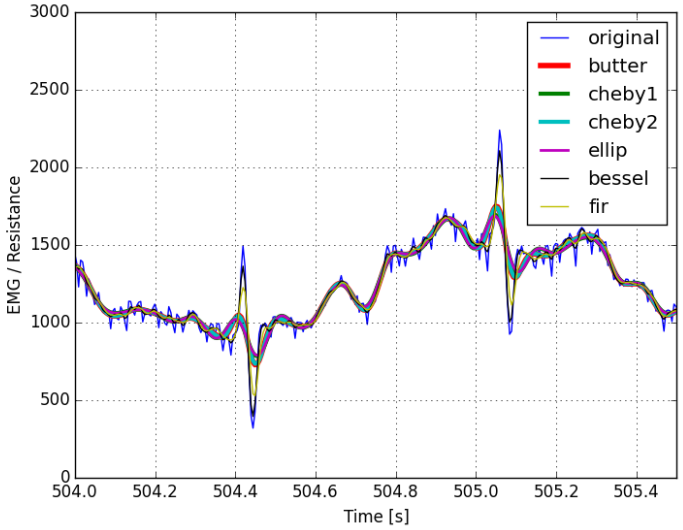

Figure 8. filter comparison for EMG raw data

4.1.2. Clustering

Including too many data points in time-series data makes it hard for users to intuitively understand a driver's behavior. k-means clustering, one of the common unsupervised learning methods, allows us to understand the general behavior at each GPS location by aggregating the locations on the track. The initial centroids for k-means clustering are determined based on the complete GPS data on track for one lap. Each centroid is picked 0.5 second intervals. Since the sampling rate of GPS is 0.1 second, clustering aggregates approximately 5 data samples within one location.

Hereby, let ${}^{l}d^{\{c,j\}}$ denote the data point within 0.1 second, where l is the lap index, c is the cluster index, and j is the data index in cluster c. Through this clustering process, the cluster index c and data index j is determined, while l is given in raw data. This ${}^{l}d^{\{c,j\}}$ owns data of both EMG data and car telemetry data with one GPS data point.

Note that this clustering only considers clean data. The classification in section 3 determines lap index l whose data meets quality levels for further clustering analysis. If the data in lap l is classified as not clean, it is simply excluded.

4.1.3. Similarity Analysis

Before computing similarity, normalization and linear interpolation are needed, as heterogeneous data points have different scales and sampling rates. First, every data point is standardized, meaning the mean equals 0 and the standard deviation equals 1.0. Second, every data point is separated with the time interval of 0.1 seconds, as we described in 4.1.2. Third, linear interpolation is applied to make the dataset comparable, because the sampling rate differs from sensor to sensor. Then, the similarity between EMG and car telemetry can be computed as

$$sim_j^{\{c\}} = \frac{1}{1 + 4\, e_j^{\{c\}}}, \quad e_j^{\{c\}} = \frac{\sum_{i=0}^{N^{\{c,j\}}} |EMG_i^{\{c,j\}} - Telemetry_i^{\{c,j\}}|}{N^{\{c,j\}}}, \tag{2}$$

where $sim_j^{\{c\}}$ denotes the similarity between EMG and car telemetry data on j th data point in cluster c, $e_j^{\{c\}}$ denotes mean absolute error between EMG and telemetry data on n th data point in cluster c, $N^{\{c,j\}}$ denotes the number of sample points after linear interpolation, and $EMG_i^{\{c,j\}}$ and $Telemetry_i^{\{c,j\}}$ and denote the i th sample points of j th data point in cluster c.

Based on multiple similarity scores in each cluster, the average and the standard deviation are computed as follows. These information is used for next step, data visualization.

$$ave^{\{c\}} = \frac{\sum_j sim_j^{\{c\}}}{N^{\{c\}}}, \quad std^{\{c\}} = \sqrt{\frac{1}{n}\sum_j \left(sim_j^{\{c\}} - ave^{\{c\}}\right)^2} \tag{3}$$

4.1.4. Data visualization for data cultivation

It is important to provide the race team the ability to cultivate data and discover actionable feedback towards performance improvements themselves. We developed a data visualization tool with a web-based user interface. This tool offers the ability to choose a parameter and instantly searching for an analytics result. Figure 9 shows the initial screen. Using this tool, users can choose the parameters of race, lap and the data analytics tool to be used. Once users choose the parameters, the result is displayed as shown in Figure 10. Here, users can PAN, zoom or perform other manipulation of the data. The examples of clustering analytics results are shown in Figure 11 and Figure 12. The meanings of the shapes and colors are described in Figure 7.

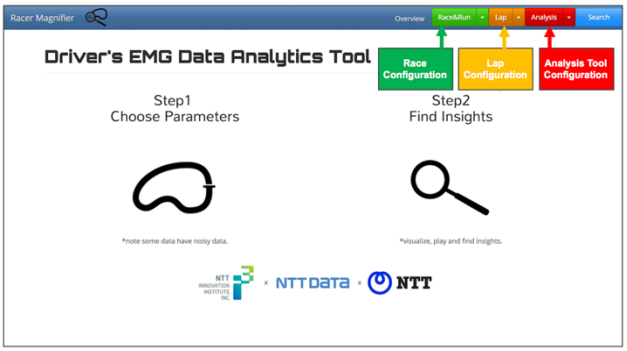

Figure 9. Parameter Selection, Race, Lap and Analytic Tool

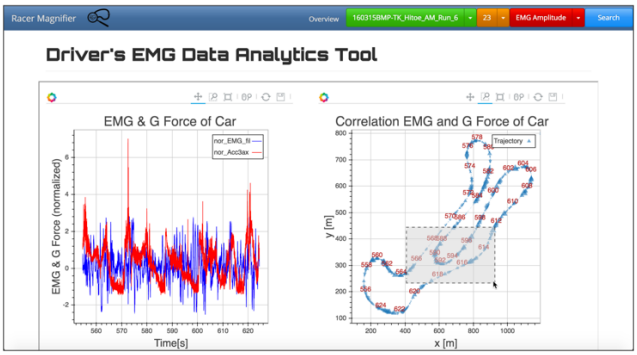

Figure 10. Demonstration of interactive tools, e.g., zooming

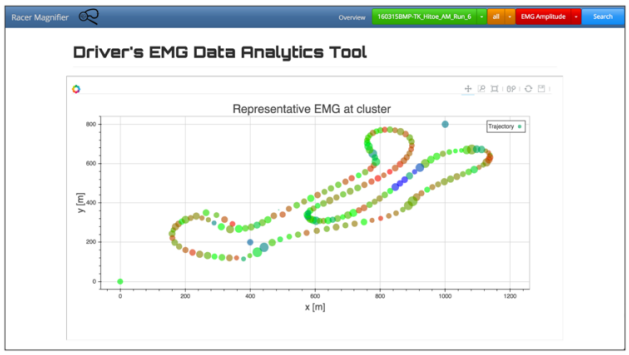

Figure 11. Example of EMG visualization

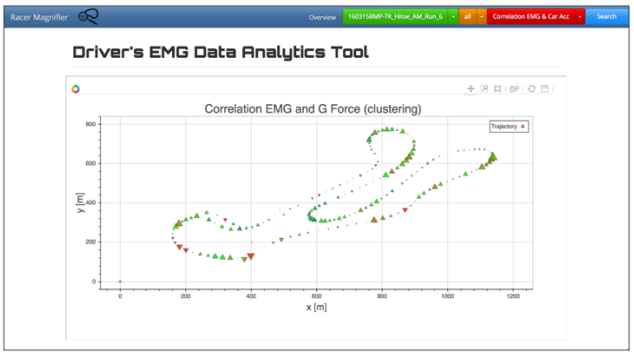

Figure 12. Example of similarity of EMG and telemetry

4.2. Evaluation – Findings from analytics

4.2.1. Findings 1 : potential improvement points towards practice focus

By using EMG, data analytics can capture the potential performance improvement points where the driver may use their muscles more consistently. In the practice laps when other competitors are not present on the track, we believe the muscle use should be consistent at each cluster within each lap. In other words, the standard deviation, $std^{\{c\}}$ in eq.(3), should be low. However, the driver may not perform well against special characteristics of the track which may be a potential improvement point.

The result of EMG cluster analysis is shown in Figure 13. For this data, we have identified the potential improvement points in red boxes. Overall, it is observed from this data that the driver tends to have high standard deviation of muscle use after turning right at a corner. This discovery lets driver identify the practice focus. In addition to this, unexpected behavior is observed in blue box area. The muscle use fluctuates on straight line, and this phenomenon happens very stably because the color is blue. This unexpected behavior could bring up another discussion on what typically happens at those specific points and how to improve performance specifically for that location.

4.2.2. Findings 2 : potential relaxation points to save muscle use

Considering the IndyCar's power steering regulation, identifying potential relaxation points is very useful for the racing team. It is expected that EMG gets higher when the accelerometer value is high because driver needs to brace his body. However, there may be avoidable muscle forces if the EMG is high when the accelerometer value is low. This behavior can be analyzed by the methodologies presented in section 4.1. When there is less similarity within a cluster, it is indicated by an inverse triangle. These results are shown in Figure 14.

For this data, we have identified the potentially avoidable muscle use as the area surrounded by red box. While the blue box may be merely the muscle use needed for steering control around the corner, the red box on relatively straight line would be the potential points to rest more. This is one example of the actionable insights discovered using heterogeneous information.

4.2.3. Subjective Feedback by professional drivers and mechanics

For this project, domain experts provided feedback as subjective evaluations. A professional race driver, Tony Kannan in NTT Data Chip Ganassi Racing said *"Those are the things the shirt has given me more knowledge about, and I've been able to re-adapt. Now I actually know how much strength I*

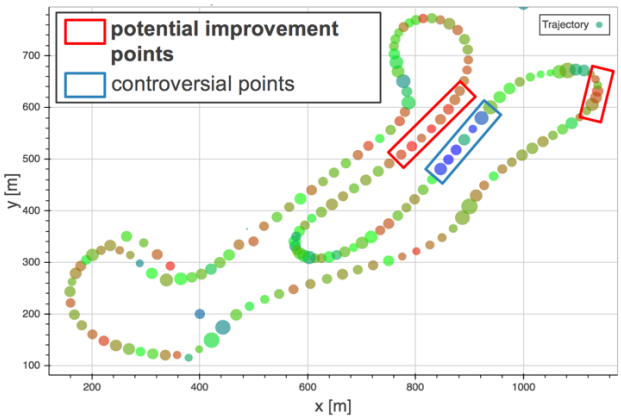

Figure 13.EMG clustering analysis

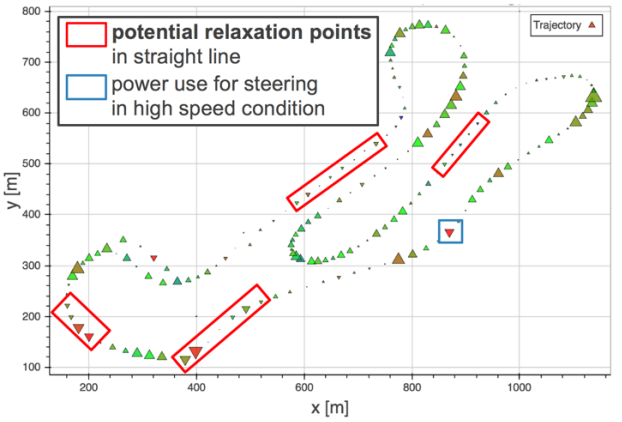

Figure 14. Similarity analysis between EMG and car telemetry

put into it and where I'm doing it. Eventually I want this thing to be able to tell my guys when I'm using too much force. Then they can tell me, 'Hey, stop squeezing the steering wheel so hard.'"

5. Conclusion

In this paper, we have tackled two major challenges: 1. the data validation of noisy signals obtained from a wearable device in extreme condition, and 2. the data cultivation to find actionable insights for the driver from multi-modal racing data. As our evaluations indicated, the combination of these two proposed methodologies have demonstrated the capability to provide actionable insights to the driver by validating noisy wearable sensor data. The methodology for the first challenge provides data quality assessment by machine learning. This technique can be applied to other use case in sports such as cycling or other disciplines of auto racing, so long as the user has heterogeneous data. This may include a real-time warning when a wearable sensor detaches from the user. The methodology for the second challenge provides a data discovery and visualization tool using unsupervised learning. One of the interesting findings is the potential relaxation points for driver to save muscle fatigue in their forearms, which is the major challenge in IndyCar. The approach of using correlation and visualizing with clustering analytics can be applied to other type of heterogeneous data analytics.

Acknowledgements

This work was done in collaboration with NTT Data Inc. who managed the data collection part in this research. Hereby, we acknowledge the effort of Mayank Gandhi and Adam L. Nelson.

References

[1] Kegelman, J. C., Harbott, L. K., & Gerdes, J. C. (2016). Insights into vehicle trajectories at the handling limits: Analysing open data from race car drivers. *Vehicle System Dynamics*.

[2] Theodosis, P. A., & Gerdes, C. J. (2012). Nonlinear Optimization of a Racing Line for an Autonomous Racecar Using Professional Driving Techniques. *ASME 2012 5th Annual Dynamic Systems and Control Conference joint with the JSME 2012 11th Motion and Vibration Conference*.

[3] Tulabandhula Theja and Rudin Cynthia (2014). Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing. *Big Data*

[4] Lee, J. H., Matsumura, K., Yamakoshi, K., Rolfe, P., Tanaka, N., Yamakoshi, Y., … Yamakoshi, T. (2013). Development of a novel Tympanic temperature monitoring system for GT car racing athletes. In *World Congress on Medical Physics and Biomedical Engineering.*

[5] Ogasawara, T., Ono, K., Matsuura, N., Yamaguchi, M., Watanabe, J., & Tsukada, S. (2015). Development of applications for a Wearable electrode embedded in inner shirt. *NTT Technical Review*, *13*