Sobriety Classification Using Gait and Other Biometrics

Kevin Martin
College of Engineering & Computer Science
Syracuse University
Los Angeles, California
kmarti44@syr.edu

Ravjot Sachdev
College of Engineering & Computer Science
Syracuse University
Syracuse, New York
rssachde@syr.edu

***Abstract*—We combine modern telemetry recording techniques and novel alcohol content recording methods with advanced machine learning algorithms to detect and identify whether or not an individual is inebriated. As the quantity of alcohol increases in the body, so too do the changes in movement and walking patterns for the individual. We supplement our analysis with similar work, ranging from other movement patterns to entirely different biometric analysis. The comparison between methods shows which direction may hold the most promise for future, real-time inebriation detection.**

***Index Terms*—BAC, biometrics, gait, sobriety, TAC**

I. Introduction

There are many ways to qualitatively and quantitatively measure a subject's blood alcohol content (BAC). Standard Field Sobriety Tests (SFSTs) might be administered by a trained individual, such as a police officer, to determine whether it is safe for a driver to continue to operate a vehicle. Breathalyzers are also used to give a precise measurement of one's BAC. In recent years, more automatic means of measuring BAC have emerged including, but not limited to, thermal eye imaging, thermal facial imaging, and gait. It is well-known that alcohol affects one's ability to walk, which is why SFSTs, such as the walk-and-turn test, measure. Therefore, we demonstrate how it is equally possible to extract features from accelerometer data to successfully analyze a subject's gait and estimate their BAC.

II. Previous Works

A. Thermal Eye Imaging

For a sober person, the iris and the sclera in the eye are similar temperatures. Therefore, thermal imaging of the eye will show little to no differences for these areas. As an individual's BAC increases, so does the temperature of their sclera. This has the effect of the iris appearing *darker* in a thermal image of an inebriated person than in a sober person, as indicated by **Figure 1**. This data gathered from cameras can be used to classify an individual as either sober or drunk [1].

B. Thermal Facial Imaging

Alcohol has a visible effect on the color of one's skin, making areas, particularly around the nose, eyes, and forehead, more distinguishable. This is due to the blood vessels becoming more active as a result of the alcohol. To other people, these areas of the face appear more *red*. These temperature differences can be measured, and a classification can be made on whether the subject is drunk or sober [2].

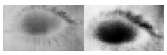

Fig. 1. Thermal images of an eye. Sober person on the left, drunk person on the right
Source: Drunk Person Screening using Eye Thermal Signatures

C. Gait and Walking Patterns

A person's ability to walk degrades as their BAC increases. This fact is what makes SFSTs, such as the walk-and-turn, so effective. "However, SFSTs were designed to make intoxication apparent to a trained law enforcement office who manually proctors them" [3]. Manual administration of these tests introduces room for error. 3-axis accelerometer data, gathered from mobile devices, makes way for a biometric system to automatically classify a user has either drunk or sober. Studies have attempted to classify participants when their BAC was both increasing and decreasing. Regression techniques used include linear regression, a Bayesian regularized neural network, and a support vector machine [Using phone sensors and...]. By comparing the results from previous studies, we hope to better determine the usefulness of our own experiment.

III. Our System

We have devised a two-part system to help identify if an individual is legally intoxicated or not. We rely first on a passive system to potentially flag an individual, then a more invasive but more accurate method to confirm. The idea is that a user won't be bothered or have any knowledge the system is working until it crosses a certain threshold at which point the user can stop and use the second phase of the system for confirmation.

The first phase comes from telemetry phone readings. The approach is based off the paper from [3], which stated that it is possible to detect if someone is or is not inebriated based on the positioning of the phone over time. We use the same

dataset to arrive at a similar conclusion. We like this approach because everyone has a cell phone, and modern cell phones are constantly recording all sorts of data including telemetry. As such, to implement this system, no changes to a user's lifestyle would need to be made.

For the original study, the alcohol content was measured using a separate device that was worn by each test user. This is very helpful for the machine learning models used later to create the link between phone movement and sobriety. However, if our system was actually put into place, the user would not need to wear an additional apparatus. This continues the theme of a very passive first stage biometric.

We felt it prudent to add a second stage to our system because, while the results are encouraging, we note that the identification and assertion of someone's inebriation is very serious and holds large consequences. Simply relying on the movement of the phone is prone to many different circumstances that could potentially alter the readings. Things like exercise, using various transportation methods, or even just general setting would all have to be controlled for to effectively roll out this system.

By using a completely different biometric, we can greatly increase the confidence in our system. We employ two different methods of data collection, inebriation detection, and control variables and if both stages arrive at the same conclusion than we feel as though it is a successful outcome. The second stage relies on a scan of the iris for the individual in question. As we discuss later, this is also a viable technique with proven results. In this paper, we don't test the effectiveness as we did with the telemetry readings, but rather leverage the work of others to confirm our implementation approach.

IV. Data Preparation

The dataset is comprised of multiple CSV files, each related to 12 individuals from a prior study. There is a separate master file which contains all the telemetry data (the x, y, and z-axis readings) measured each nanosecond, classified by individual [4]. This file contains approximately 14 million telemetry readings. There are also 12 files (one for each participant) which hold the alcohol readings throughout each time frame. The alcohol is measured in TAC (Transdermal Alcohol Content) which allows for less invasive data collection than something like a traditional breathalyzer. Additionally, the telemetry readings are taken every nanosecond, whereas the TAC readings are taken every thirty minutes. As such, there are significantly less TAC reading records than there are telemetry readings. Finally, we note the distinction between BAC and TAC. In general, the two are very similar in terms of reliability and differ only in their means of collection. Importantly, BAC would be considered an active biometric while TAC is passive. Thus we will rely on TAC exclusively.

A. Feature Extraction

The master file, which holds the telemetry readings, was the starting point. We imported the file using the Python library Pandas and created a dataframe. This allowed us to easily manipulate and match up the participants and their readings. We created a new column that was a copy of the original timestamp column, only this time it was formatted in seconds. This allowed for us to match the TAC readings (which are formatted in seconds) to the accelerometer readings (which were originally formatted in milliseconds).

Next we created our new features. Using a single point in time (x, y, and z positioning) could not possibly help identify a person's inebriation level. We are focusing on motion. In order to capture motion, we used three rolling averages for each of the three axes. We used a rolling previous 500, 1,000, and 5,000 mean to approximate motion volatility. Our thought was that the higher the rolling averages were, the more likely the individual was to be over the legal limit. See **Figure 2** for an example of how rolling averages and TAC display similar movements throughout the relevant time period. Additionally, we kept all three axes separate, but used all as features in our classification algorithms. This allowed us to test three different features for each TAC reading across three different time frames, for a total of nine unique features.

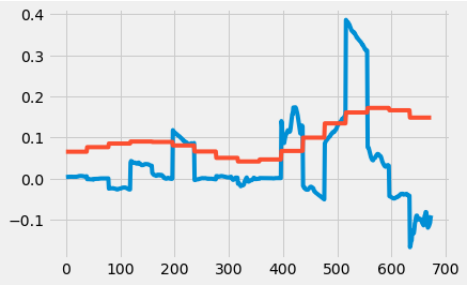

Fig. 2. Sample rolling average of the x axis vs. TAC reading

B. Data Manipulation

Next we imported each user TAC reading data, which were saved as separate CSV files. Again we used Pandas to create a dataframe for each, only this time we added a merge function to also pull in the accelerometer readings. Because the TAC readings were done every 30 minutes, we were only able to match a few thousand user timestamps. However, this also meant we could include all of the nine rolling averages as well. In effect, we were pulling in not only the absolute x, y, and z readings, but also the previous 500, 1,000, and 5,000 averages. Our rationale was that since the TAC readings were so spread out, the accelerometer readings between those measurements were largely irrelevant. Thus we focus on only the critical times.

After merging the accelerometer readings to the user data, we had 12 new dataframes. We kept these preserved, but found it most useful to combine all into one summary dataframe using the simple concat function built into Pandas. This left us with over 5,000 instances of TAC readings matched to precise accelerometer readings as well as our nine rolling averages. From there we could extract both user-level data as well as aggregate information.

V. Methodology

We had two main goals for the actual testing. First we wanted to see if we could, based on our rolling averages, predict the TAC level of a participant. This takes the form of a regression problem as the target variable is continuous. We look to use multiple methods and evaluate the results of each: simple linear regression (as a baseline), a decision tree regressor, and a random forest regressor.

A. Regression Algorithms

Linear regression attempts to fit a line that minimizes some cost function based on errors, with the key assumption that the relationship between the input attributes and the target is linear. It is the least sophisticated algorithm, and may or may not be considered a true machine learning approach. That is not to say it is without merit, and indeed we appreciate how quick and easy it is to implement. We use this as the bare minimum: any other algorithm should be able to beat this one, and if not, some further investigation would be necessary.

Next we implemented a decision tree regression algorithm, which works by splitting the dataset into separate nodes. Each node is intended to be similar to one another thus attempting to reduce impurity. Once we reach the final leaf nodes, the predicted value for that leaf will be the average TAC reading of all entires in that leaf. To implement, we utilized the package scikit-learn, which includes a decision tree regression model.

Finally, we looked a random forest regression algorithm. This idea builds on the previous decision tree, but actually constructs multiple decision trees (forest) and the output is the mean value of each of these separate decision trees. As this the most sophisticated of the three approaches, we would expect this to be the top performer.

B. Binary Methods

Our second approach, and the one that we will focus on more for our biometric system, is a binary classification system. Our goal is to predict whether or not someone is over the legal limit, so while the regression results would be interesting, they would not prove as useful for our overall goal. Additionally, we believe that it is an easier problem to solve in that the classifier can be more imprecise with its predictions.

In order to train classifier algorithms, we added a new column to the summary dataframe. The column would be populated with a "1" for any TAC greater than or equal to .08, and a "0" for any reading less than .08. This new column will be the target column for which we will have the classification algorithms focus. This gives us a decent distribution, with about twice as many sober readings as inebriated, which can be seen in **Figure 3**.

Again, we utilize several different methods, each varying in sophistication, and compare the results. First we use a Naïve Bayes as a general baseline. Similar to linear regression in terms of its ease of implementation and relatively low computational complexity, we use this as a quick check to ensure the more robust methods are outperforming this one.

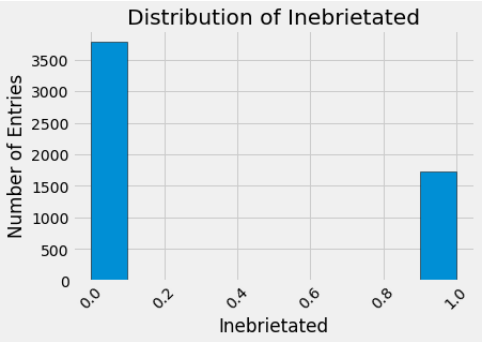

Fig. 3. Binary classification results

Next we try a decision tree classifier without any hyperparameter adjustments. This also serves as a baseline, but more for the tuning aspect. To (hopefully) improve our performance, we try a grid search, which uses cross-validation to help find the optimal hypterparameters for the decision tree. By creating a pipeline of inputs to test ahead of time, we allow the grid search to test all possible combinations and output the results. We consider the decision tree and the grid search to be two different methods even though they rely on the same underlying algorithm. This is another situation where we have high expectations about the outcome (in this case that the grid search should, in some capacity, outperform the basic decision tree).

Finally, we use a "pre-tuned" model, XGBoost. XGBoost is a variant on gradient boosting, and generally performs very well. This algorithm benefits from a few key things, namely the ease of deployment as well as the freedom from heavy hyperparameter tuning. This model is as quick/easy to deploy as any of the above, and in fact far easier than the grid search, yet consistently provides exceptional results. We expect this to be one of the best models for the classification section.

VI. Results

A. Regression Testing

Initially, we used three separate regression models to analyze the correlation between the moving averages from the telemetry data and the TAC readings. The three regression models we chose were linear regression, decision tree regression, and random forest. For each regression model, we calculated the root mean square error (RMSE) and the mean absolute error. Of the three models, the random forest model performed the best, with a RMSE of approximately 0.0085. In comparison, the RMSE of the linear regression model for the same dataset was approximately 0.049, and 0.011 for the decision tree.

While useful to compare against each other, we are not particularly confident with these measurements. We do not believe that any thresholds have been crossed to make use of a regression technique for our biometric system purposes. The importance of the problem is too great to rely on these types of benchmarks in an objective fashion. We are able to determine,

between the three used here, which would be the ideal choice, but still wish to focus on the binary classification.

B. Binary Classification

To examine the data through a more concrete lens, we added a field to group the TAC readings into one of two classes: inebriated or sober. The dividing line was a TAC reading of greater than or equal .08, the legal limit for operating a motor vehicle in most states. This allowed us to use the binary classification algorithms, and the results were very encouraging. Now we can objectively see not only the accuracy or performance of the model, but also describe it in legal terms.

Based on the three sets of rolling averages for each axis reading, three of our algorithms are able to achieve accuracy ratings of over 97%. The Naïve Bayes grossly underperformed every other algorithm. We believe that the results generated are wholly unreliable and would discard its output completely in a real-world setting.

Next, as expected, was the original decision tree. Without any hyperparameters tuned, we would expect it to exhibit less accurate performance. However, after running the grid search, we see only the modest of gains, with a nearly imperceptible increase in accuracy, recall, precision, and F-Score (F1) across the board. While we are pleased to see gains of some amount, the result is disappointing given the pipeline infrastructure needed to validate the grid search. Furthermore, it is much more computationally expensive to run a grid search, yet yields almost no improvement over the regular decision tree. Thus we would not recommend the grid search either, and recommend staying with a decision tree.

Our most successful method, XGBoost, which had an accuracy of over 98%, was able to beat not only the decision tree, but also the grid search with optimized hyperparameters. The set up and train time for this approach was almost the same as the regular decision tree, but yielded over a full percentage point more in accuracy. The ancillary metrics also benefited as well. This is the clear choice, both in terms of cost but as well as benefit.See **Table 1** below for a summary of these results:

TABLE I
SUMMARY OF BINARY CLASSIFICATION METHODS

Method	Hyperparameters	Accuracy	Precision	Recall	F1
XGBoost	N/A	98.97%	97.84%	98.81%	98.32%
Grid Search	criterion, depth	97.76%	95.34%	97.42%	96.37%
Decision Tree	None	97.70%	95.33%	97.22%	96.27%
Naïve Bayes	None	35.69%	30.65%	87.89%	45.46%

VII. HOW OUR SYSTEM COMPARES

Classifying an individual as sober or inebriated using a biometrics system is hardly a novel idea. In fact, there have even been studies that used gait. Comparing the results of our experiment with previous studies helps to determine which methodology may prove more accurate or more beneficial for future automatic detection of inebriated individuals.

A. Facial Thermal Imaging Techniques

Normally, the temperature of a person's skin is around 33.5°C, but after consuming alcohol "there is a temperature increase in capillary density, such as around the nose, forehead, and eyes" [4]. This provides measurable features that could be used by a biometric system to determine whether a subject is inebriated.

One study collected thermal images of participants and attempted to determine how many beers they had consumed, ranging from zero to four beers. As seen in **Figure 4**, a 22-point grid was overlaid over each face to obtain a 22 dimension feature vector. A Gaussian Mixture Model (GMM), a supervised learning algorithm which utilizes Gaussian distributions to form a probability density function, was selected to classify the images into the five different classes.

The test data selected for the study was from the class of subjects who drank 4 beers. In the end, the biometric system classified approximately 13% of these subjects as *sober*, meaning it had an accuracy rate of about 87%. One of the limitations with this study is in the test data, as the test data was limited to non-sober participants only.

Another study followed a very similar methodology: a 22 dimensional vector used as input to a GMM algorithm. In this study, instead of trying to predict the number of beers one had consumed, they simplified it to a binary classification problem of either drunk or sober. Although the problem is simplified, it is important to note that the results for this study seem misleading: "When the classification was done on drunk drivers approximately 13% were classified as drunk and 87% as sober. Hence the accuracy of the classification stage is 87% approximately" [5]. Even if the accuracy was 87%, as the paper states, our system outperforms both of these systems that use thermal facial imaging.

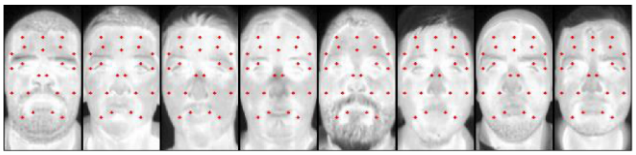

Fig. 4. 22-point grid overlaid over thermal images

B. Eye Thermal Imaging Techniques

Iris identification is hardly a new biometric, yet the eyes contain other features that are usable for autonomous classification. By focusing on temperature changes in the iris and sclera, researchers have proposed that a system could accurately determine whether a user was inebriated. Data collection would be more akin to the facial thermal imaging technique, naturally, as both require use of a thermal camera, whereas our gait analysis would not.

One particular study of eye thermal imaging analyzed the effect of alcohol on 41 individuals. Each person was given four glasses of red wine within an hour, and 50 frames of images were collected at two separate instances: before any alcohol

was consumed and less than an hour after the fourth glass of wine. The breathalyzer measurements produced a fairly wide from, with a minimum of 0.22mg/L and a maximum of 0.9mg/L. In BAC, this range would be approximately equal to 0.055%BAC to 0.225%BAC. In 28 of the 41 participants, nearly 70%, no image processing was necessary to view the darker, indicating cooler, iris in the thermal images of drunk subjects, when compared to images of the same participant sober. For 8 of the remaining 12, some form of histogram modification algorithm, to stretch a designated range of the histogram, was necessary to view the expected changes. Finally, for each of the final five participants, no temperature changes were detected when comparing the thermal images taken when they were drunk to those taken when they were sober.

Using these images, features can be extracted using the mean value of pixels in the iris and the sclera. One possible feature is the mean value of the pixels in the sclera divided by the mean value of those pixels in the iris region. Another feature is is the variance of the value of the pixels in the entire thermal image of the eye. In both scenarios, the study determined that a classification can be made whether a given individual is drunk or sober with 99% confidence [6]. This performs slightly better than our best method, that which uses XGBoost, and offers another possible avenue for classifying drunk individuals.

C. Gait Techniques

To critically analyze the practicality of our proposed biometric system, it would be quite useful to compare it to a similar system, specifically one that uses gait analysis to classify drunk individuals. One such study also utilized accelerometer and gyroscope data gathered mobile devices, but also mixed in compass data, whereas our system does not. In this particular study, subjects were given 3-4 ounces of alcohol approximately every 15 minutes until their BAC was measured to be 0.1. The BAC was then measured every seven minutes until the alcohol was determined to no longer have a significant impact on their motor skills. Throughout the study, subjects were asked to perform a few SFSTs, to gather accelerometer data. Features included, but were not limited to, postural sway features and zero crossing rate, and were using to train a a few different models, such as XGBoost and convolutional neural networks. The RMSE was then measured to compare each model's performance. Of the models selected, the Long Short-term Memory (LSTM) networks performed the best, which is not a technique that we used. One of the major conclusions of thisepaper is that the RMSE for the various models was generally lower for descending limb data. That is, the systems performed better after subjects had stopped drinking [3]. This is an important distinction to our own study which does not make any such differentiation. This introduces another possibility for improving our own system but improving the input data and analyzing gait features on different limbs of the drinking process.

Another technique for using gait analysis does not use accelerometer data at all. In fact, it simply uses a video sequence of a person walking. From here, the individual frames were converted into optical flow images to view the velocities of the subject's movements, as seen in **Figure 5**. Features extracted from these frames were than passed into an SVM classifier, The study claims a near 100% accuracy of identifying drunk individuals using gait analysis [2]. While impressive, we have a few hesitations regarding the study. For one, the dataset is limited, using videos of just four students to train the classifier. The bigger worry is how the data was gathered. For the sober class, students walked normally. Yet, for the drunk class, students *imitated* a drunk person's movement. The accuracy of their imitations would necessarily affect the training data, and with such a limited dataset, the near 100% accuracy is something worth questioning. At least the study was not entirely on gait analysis, as it combined it with both eye and face thermal imaging. Nonetheless, using video images rather than accelerometer data is an interesting proposition, and certainly something to consider in the future. One of the advantages is that it does not require access to data from the subject's phone. If a special app is required to gather the appropriate accelerometer and gyroscope data, as in the previous study, then camera data provides an alternative for autonomous classification in situations where it is not guaranteed that the user will have the appropriate app installed.

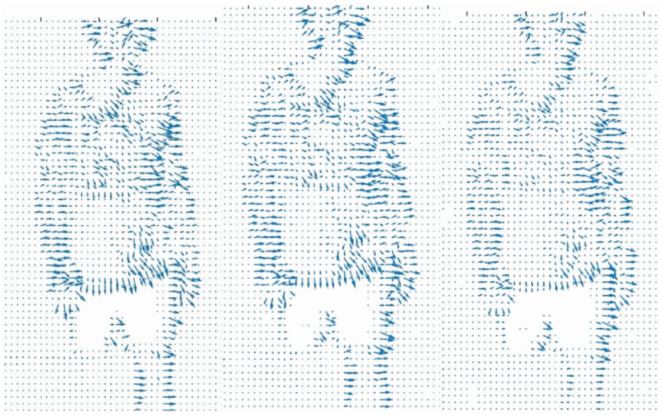

Fig. 5. Optical flow frames

VIII. Future work

Our proposed system consists of a two-step process: an initial classification made using gait-based analysis, which prompts for an iris scan on a positive classification. The reason for behind this decision is two-fold. First, although we were able to create a classifier using gait measurements with fairly high accuracy, our dataset is fairly limited, with only 12 participants. Second, we've seen through other studies that iris scanning is slightly more accurate than our gait analysis. Therefore, we'd feel more confident in our classification if we were to combine these two results. One of the disadvantages of this is that iris thermal imaging analysis is not readily

available in a mobile device, whereas accelerometer data is easily obtained. So a new hardware must be added to the mobile device or a second device entirely would be needed to allow for this second stage of analysis.

A single stage analysis may be sufficient in the future if further studies are done. A larger number of subjects would be the first step to providing more reliable results through just gait analysis. It would also be important that this data set is expansive. That is to say, data is gathered from a group of subjects representing a wide range of ages and drinking backgrounds. It has been shown that alcohol affects the BAC of people differently depending on how much they drink, weight, ages, gender, etc. As such, the data must represent this. Future gait analysis could also involve video data, should iris imaging be deemed cumbersome.

IX. Conclusion

In this study, we investigate various methods to automatically classify an inebriated individual. We use a data set of telemetry readings from 12 individuals to build our own classifier focused on gait analysis. In combination, we propose a second stage to the biometric system which would utilize thermal iris images. We briefly discuss a few previous studies revolving around classification of drunk individuals and compare our techniques to those that existed previously. It is our hope that as more research is done on these methodologies, a more accurate and more secure system can be created to utilize readily available biometric data to classify individuals in a non-invasive manner.

References

[1] Koukiou, G.; Anastassopoulos, V. Drunk person screening using eye thermal signatures. J. Forensic Sci. 2016, 61, 259–264.

[2] M. K. Bhuyan, S. Dhawle, P. Sasmal and G. Koukiou, "Intoxicated Person Identification Using Thermal Infrared Images and Gait," 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2018, pp. 1-3, doi: 10.1109/WiSPNET.2018.8538761.

[3] R. Li et al., "On Smartphone Sensability of Bi-Phasic User Intoxication Levels from Diverse Walk Types in Standardized Field Sobriety Tests," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 3279-3285, doi: 10.1109/EMBC.2019.8857214.

[4] Killian, J.A., Passino, K.M., Nandi, A., Madden, D.R. and Clapp, J., Learning to Detect Heavy Drinking Episodes Using Smartphone Accelerometer Data. In Proceedings of the 4th International Workshop on Knowledge Discovery in Healthcare Data co-located with the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019) (pp. 35-42).

[5] Hermosilla, G., Verdugo, J. L., Farias, G., Esteban, V., Pizarro, F., and Machuca, M. (2018). Face Recognition and Drunk Classification Using Infrared Face Images. Journal of Sensors, 2018, 8. doi:http://dx.doi.org.libezproxy2.syr.edu/10.1155/2018/5813514

[6] S. Menon, S. J., A. S.K., A. P. Nair and S. S., "Driver Face Recognition and Sober Drunk Classification using Thermal Images," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0400-0404, doi: 10.1109/ICCSP.2019.8697908.

[7] Koukiou, G.; Anastassopoulos, V. Drunk person screening using eye thermal signatures. J. Forensic Sci. 2016, 61, 259–264.