Annotation and Position Recall from Low Grade Sensorial Data in the Context of Topological Railway Maps

Vlad Doru COLCERIU, Teodor STEFANUT, Victor BACU, Dorian GORGAN*

Computer Science Department, Technical University of Cluj-Napoca

28 Memorandumului, Cluj-Napoca, 400114, Romania

vlad.colceriu@cs.utcluj.ro, teodor.stefanut@cs.utcluj.ro, victor.bacu@cs.utcluj.ro, dorian.gorgan@cs.utcluj.ro (**Corresponding author*)

Abstract: Describing high quality track maps is an important prerequisite for the localization algorithms, which depend on relatively unreliable and unsafe triangulation technologies, such as the Global Navigation Satellite System solutions available on modern mobile devices. The state of the art studies and experiments concerning the maps annotation methodologies have been focused on using large quantities of crowdsourcing sensorial data and open source topological maps. The objective is to produce a low-cost platform for regional railway operators, to visualize the track's geometry and to generate a track representation, which can be then recalled across multiple journeys. The paper aims to develop and experiment with a sensorial signature representation toward providing a flexible solution for signal alignment based on Genetic Algorithms. The alignment solution intends to describe the sensorial signals of the track line in a relevant manner that makes localization algorithms possible based on the nonlinear state estimation.

Keywords: Rail transportation, Crowdsourcing, Open source data, Signal processing, Sensorial data patterns, Nonlinear state estimation, Localization algorithm.

1. Introduction

Railway positioning applications are safety critical in nature and require a proven safety level. The problem can further be subdivided into multiple sub-problems, ranging from precise along track positioning to train integrity detection. Such applications concern multiple heterogeneous track and train side elements and are made up of hardware and software components, which are required to have a proven safety level above the seven 9 range (99.99999\%).

Hardware and software vendors must, as a result, offer safety and security assurance for the products they supply, since they are being held liable for any damages or loss of human life being a consequence of using their solutions. This has led to de-facto standards, such as IEC defined Safety Integrity Levels to be implemented across the domain.

Standardization attempts have lead thus far to the development of applications and procedures offering the required reliability, safety and security, such as the ETCS/ERTMS [15] in Europe and PTC [16] in the US. But what these solutions lack to address is financial accessibility. Many regional train operators do not have the money or resources required to undertake the implementation of such solutions, which require both train and track side components, to be installed and maintained. As is the case of the PTC system, which was mandatory to be implemented until 2015, for all railway providers, but has still not managed to gain wide range of acceptance among regional operators.

To better understand the complexity of train positioning on the railway, the problem has been conveniently divided up into five disjoint subcategories: (1) detection of along track position; (2) detection of the current track in a multiple track environment; (3) train speed and braking estimation; (4) detection of turnout direction and (5) train integrity detection.

The current work proposes the use of GPS-based solutions combined with IMU (Inertial Measurement Units) to solve the detection of along track and parallel track position. Unfortunately, the problem with both solutions is that, even when combined, they do not provide sufficiently good results.

First of all, commercial GPS satellite navigation systems are not precise enough to be able to distinguish between adjacent tracks in a multi-track environment, which can be as close as 4 meters apart [22], while the 98'th percentile of GPS accuracy falls in the range of 16-20 m [14].

IMU sensors on the other hand have been proven to provide an significant improvements and even constitute better alternatives to pure GNSS based solutions by augmenting the positioning stream with information about kinematic state, as has

https://doi.org/10.24846/v26i4y201711

been extensively explored in a number of works, treating everything from turnout crossing direction [26], [8] and train parting [1], [24] detection and even train speed estimation [18].

This work proposes to use a hybrid approach to localization by augmenting sensor data with position, instead of augmenting position data with sensor data. Through this approach, we are generating uniquely identifiable sensor characteristics as presented in a similar work [17], which then can be used to inversely detect position through matching sensor characteristics. While this is a very generous promise, the work presented goes so far as to prove the fact that a repeatable signal can be identified by using multiple train journeys.

The paper is structured as follows. Section 2 discusses some of the other achievements in the field of train localization and presents some similar applications from the train and auto industry. Later in section 3 a discussion of previous research about position improvement and signal noise reduction is presented. Only in sections 4 and 5 do we discuss the mechanisms for signal pattern modeling and signal shift adjustment. The signal shift adjustment is an operation designed to adjust the differences observed between signals, collected during different train journeys, to obtain a correlated signal database, to be later used for position detection. Finally, in section 6 the conclusions and future directions of research are presented.

2. Related Works

A straight forward way to approach train localization is by using satellite based global navigation. As a result, in literature there are many solutions promising to solve localization using such an approach. Out of these, of interest are two solutions, mainly TrainGuard [25] and Satloc [4], both of which approach localization by using high-end EGNOS (European Geostationary Navigation Overlay System) receivers to improve the accuracy of the GPS, by removing atmospheric effects on the signal. Although not enough to offer the desired safety levels, these applications take another route to assuring safety, by modeling ahead of time the worst-case scenario for satellite constellations and by assuring that no obstructions may exist in the Line-of-sight of the receiver. Additionally, the Satloc system uses a wheel counter to obtain a better estimation of speed and thus allows for a more hybrid approach to localization.

Although unaided GPS/GNSS localization although might suffice for an advisory system in a controlled environment, there is no hope of satisfying the safety and security constraints imposed by the railway domain, since there are a series of factors, such as multi-path interference, sources of intentional and unintentional jamming, signal delays and other factors, which make reliable positioning impossible [27].

A series of approaches exploring the possibility of using IMU, such as accelerometers and gyroscopes to detect train positioning have been explored in papers [5], [8]. Additionally, the idea of using acceleration and speed of train to calculate train location has been extensively researched in [18], [1] and by Scholten in [24], which also discussed the issue of train parting detection.

Vehicle vibration has also been extensively discussed in literature and used in diverse domains, such as terrain identification in planetary exploration rovers [7], road quality mapping used in bikes [23] and cars [9], [20] and finally in mobile based applications [2].

Furthermore, some experiments with using multiple GPS constellations to detect vehicle position have been discussed in [28]. The application presented, Tiramisu, also offers bus-ridership information using a crowd outsourcing approach.

3. Signal Collection and Processing

The current paper presents a brief overview of the signal collection and processing techniques implemented to obtain the fundamental signals used in pattern modeling and recognition. The algorithms and results presented have also been discussed in a previous paper [12], treating the annotation of topological railway maps with sensor data. While the previous paper focused on the advantages of the signal collection and signal processing solutions, the current paper focuses on their shortcomings and makes an argument for extending the algorithms to include the possibility of automated signal recognition.

Before delving into the localization issues and signal filtering approaches it is worthwhile briefly discussing another concept presented in a previous work of the authors [11]: a novel conceptual methodology of obtaining a track description by using only MEMS sensors, such as accelerometer and gyroscope. These are used to create representations of the track environment by mapping their signal values to their along track location.

Signals and positions are intrinsically linked in modern mobile devices. Most of them offer access to GPS/GLONASS position to provide location sensitive services to its users and have at least accelerometer and gyroscope to improve device usability, by detecting contextual gestures, such as screen rotation.

To take advantage of this fact, the paper [10] proposes the usage of a crowd sourced solution to integrate multiple devices measuring the same contextual environment to describe the railway track. By opportunistically scavenging the data collected by devices belonging to the riders of a train, one may measure the movements of the train.

Due to the inherent one-dimensional nature of the railway environment and small number of variables, used to describe train movement one may easily translate the movement of the rail cart to general characteristics of the track. Although there are other variables which may affect train movement (such as cart damper characteristics, position of the recording sensors inside the cart and even temperature conditions affecting the metal beams in the track), these are not considered in this study. As a result, the only remaining variable to be considered is the train speed, which affects both accelerometer and gyroscope measurements.

3.1 MAX-AV Localization Algorithm

In paper [11] we proposed an algorithm to collect gyroscopic and acceleration data from multiple devices placed in the same rail cart, across multiple journeys on the same line, in order to generate a uniform representation of the track at a given point. In some respects, the mentioned algorithm has succeeded in generating a visually appealing result, which may be used in line annotation for the benefit of human users, but not for an algorithm hoping to recognize the signal over multiple journeys and generate a unique representation.

As presented in Figure 1 the developed algorithm has three sequentially dependent phases: sensor data collection, MAX-AV position improvement and signal denoising based on Multi Resolution Wavelet Analysis.

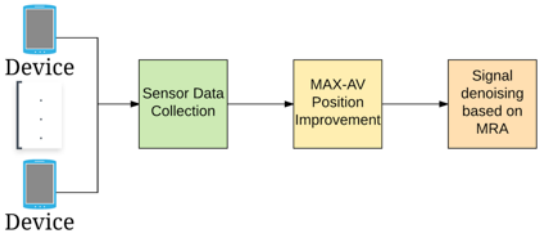

Figure 1. Pipeline of track annotation algorithm

The first step of the algorithm establishes a way in which all signals belonging to the same journey are aligned. The GPS time, which due to the nature of its application must be very precise, was used to align the signals coming from different devices. This fact is very important since there is no other reliable way to establish a global absolute time.

Once a common time frame was established, the second part of the algorithm could start. Positions from multiple devices riding on the same train cart could be fused together using the localization improvement algorithm, named MAX-AV. MAX-AVerage, as the name suggests, is an optimized averaging algorithm, which uses DBSCAN to calculate the average of the closest cluster to the railway line. As expected, the parameters of the DBSCAN algorithm require an iterative optimization to achieve optimal results. This step is done by a gradient descent algorithm which tries to reduce the localization error by optimizing the ε-value (expected distance between members of a cluster) and minimum number of points in a cluster.

What makes this algorithm different is the way it calculates positioning accuracy, because to calculate it one needs both lateral track accuracy and along track accuracy.

Since there is no way to know the exact along track position, only the lateral distance from

the track line, a calculation of along track position must be performed. By subtracting the geometric projection of the estimated position of the measuring device from the point the device should have arrived given its reported speed and previous location, one may come to a reasonable estimate. Although measureable and precise, it is still subject to any speed imprecisions.

The results of the MAX-AV algorithm were reasonably accurate, offering an average improvement in the variation (σ^2) from the optimal track position of 61.27 m^2, when compared to the closest journey. With all this improvement, there was no way to tell the exact position of the train, since the results were heavily dependent on the velocity noise exhibited by the GPS sensor and thus any signals compared across multiple journeys would suffer from alignment issues, which will be further discussed in section 5.

3.2 Wavelet based Noise Reduction Algorithm

The last part of the algorithm for generating the base signal is the noise filtering since the signal correlation for each axis of a sensors, was in the 0.0 to 0.1 range, thus very small, measures had to be taken to correct for the noise.

A wavelet decomposition algorithm is used to calculate the frequencies, which were least affected by noise. By using the Multi-Resolution Analysis wavelet decomposition algorithm [12] [13] we could decompose the signal into a total of 12 dyadic frequencies ranging from 128 Hz to as low as 1/32 Hz. Then, by analyzing the correlation of the signals of the devices from a single journey, we could extract only the signals having high correlation (> 0.7).

The result of running this algorithm yielded consistently the same output for all journeys. Acceleration on the X-axis (lateral to track) and Y-axis (longitudinal to track), as seen in Figure 2, showed consistently high correlation (> 0.7) in the frequency bands of 1/2 to 1/32 Hz. The same was true for the Y-axis(roll) and Z-axis(yaw) of the gyroscope, but in a much smaller frequency range, limited to 1/32 Hz.

Despite such good results for signal correlation inside a journey, signals collected in different journeys, when compared, did not show such good correlation. These signals, although outputting sufficiently good correlation to be interpreted by a human, did not fare well under an automatic analysis. This is mainly due to unreliable shifts in signal position caused by the limited accuracy of the GPS positioning.

4. Signal Pattern Representation

This section presents the research done to unveil a solution suited to condensed signal representation, allowing the signal transfer from a computationally capable backend to a resource scarce mobile device frontend. By allowing the backend to run intensive signal processing algorithms, we limited the task of the frontend to the recognition of track patterns.

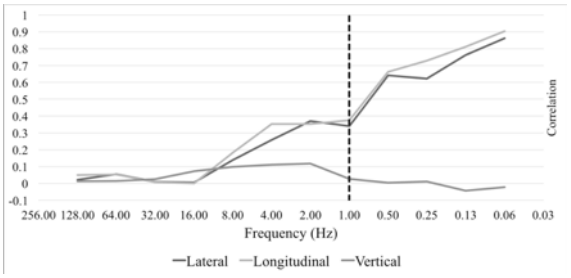

Figure 2. Average correlation of acceleration signal across axes and component frequencies

The signal had to be condensed in such a manner that it could quickly be transferred from permanent storage to a volatile mobile device storage in a matter of minutes, through a low bandwidth connection. Using our approach, a signal which requires in the order of 1 GB of storage on the server-side has been reduced to just a few MB.

The presented use-case pertains to the transfer of features ahead of time for a given station to station route. The features must be transferred in the order of minutes, as required by train wait times for allowing of boarding, as presented in the work [10].

Using brute force recognition techniques, such as cross-correlation, which has the advantage of being shift invariant, is not enough to solve the issue of small variations coming from the fact that the sensors are not strapped down to the train chassis during collection and that the sensor network is heterogeneous and that each train has its specific wheel damper model, etc.

Instead, this work focuses on edit-distance based string comparison techniques [19], [3] to be used in both signal recall and later in a future work in online Particle Filter signal recognition.

So, instead of concentrating on exhaustively modeling all factors affecting the signal, focus was put on detecting easily reproducible patterns which could be represented as string characters, which in their turn could then be easily recognized by using an edit-distance approach.

When starting to implement the pattern recall algorithm it was decided not to make any assumptions about the shape or size of the signal to detect the minimum granularity of a character string. The idea was that the smallest building blocks of a signal were unknown and thus might be composed of multiple overlapping basic signal features, which we called Camel Humps, as described by Figure 4. These signals would form a database of expected signal shapes and assigned a character from our character-set.

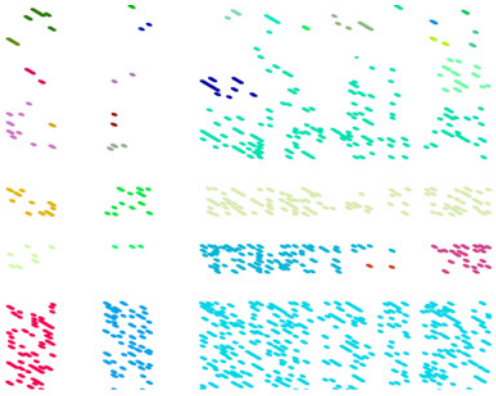

Figure 3. Journey subset containing clustered self-detecting patterns resulting from cross-correlation of original signal

4.1 Detecting Repeating Signals using Image Processing Algorithms

The first unsuccessful approach involved using repeating patterns to detect feature subsets in the signal. To this end, inspiration was derived from the music domain, where pattern recognition is used for retrieval and indexing of music tracks, browsing into a song and audio thumb nailing. Because of this, the research of Aucouturier [3] was found to be promising in solving the problem of signal compression by detecting most common patterns and encoding them into a character database.

The algorithm worked by transforming the one dimensional musical signal into a two-dimensional auto-correlation image, in which each point (i, j) of the image represented the correlation of the i-th and j-th point in the signal, where both point i and j were encoded as integer values. This could be done by collapsing the signal into the learned polyphonic timbres of different instruments.

After removing trivial occurrences of the main diagonal and upper echelon sub-patterns, a Hough transform algorithm [3] was applied to discover distorted continuous lines from the ideal 45° angle and thus account for edit-distance operations, such as deletion, insertion and substitution.

The limitation of the above presented algorithm showed itself when dealing with a continuous signal exhibiting values in the [-2,2] range. The problem was that the auto-correlation generated continuous and fuzzy values in the [0,1] range, instead of being crisp 0 and 1 values. This made the algorithm less stable, since a threshold for correlation had to be found to make the image segment correctly.

This lead to an unstable diagonal resolution algorithm, which even when modified to finding optimal clusters of segments to detect bounding squares or trivial auto-correlation, did not show good results. An example of the results obtained by this algorithm can be found in Figure 3, where the region of a detected cluster is highlighted in the same color. The suspected reason for the lack of results is the fact that the algorithm as it stood required a dynamic threshold for different regions of the signal, so that trivial correlation rectangles could be formed. More about the subtleties of this algorithm can be found in the work of Aucouturier [3].

4.2. Encoding Data into Humps

As a result, a closer look was taken at noise filtering algorithm and the dyadic frequency components it generated. The similarity between the obtained signals and the original Daubechies wavelet [13] used for noise filtering was observed. The relation was to be expected due to the orthogonal nature of the transform, which decomposes the signal into scaled copies of the mother wavelet (Figure 4).

Starting from this point of similarity, a set of characteristics were devised to uniquely identify each component of the signal. In this regard, the smallest signal component was taken as benchmark for the signal since it can easily be divided into 3 simple yet descriptive elements: *position(p)*, *width(w)* or *length* and *magnitude(m)* or *peak*.

Given that the signal marker representation contains just 3 variables: (p, w, m), the signal could easily be processed by using the above described structure, for both identification and recall. However, the problem of a deterministic edit-distance based approach is limited by the prohibitive computation time. As presented in paper [6], the computational complexity of an ordered edit-distance algorithm, when using an efficient tree-structure, is shown to be $O(|T_1||T_2(I_1+I_2)^2)$. Here, $|T_1|$ is described as the size of one of the trees and I_1 is the maximum in-degree of the tree.

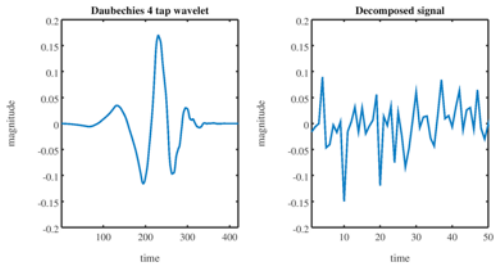

Figure 4. Signal comparison between Daubechies 4 tap wavelet and observed signal

Because the minimal unit of the signal is the Hump, the amount of POI (points of interest) is considerably reduced, since the width of the selected pattern even in its densest cases is not less than 3 m. At the same time, the minimal position resolution of the signal, as it has been selected, stands at 0,125 m.

It is also worth noting that, for any kind of relevant transformation to occur from time to along track distance, the train must always experience movement for the signal to have any kind of relevance. Therefore, all time intervals for which train speed has fallen below 2 *m/s*, have been cut out of the signal. More about the reasoning for keeping a minimal speed of transformation is explained in [10] and [11].

By using Camel Humps (p, w, m) to represent patterns, the signal size has been reduced from values in the 0.5 x 10^6 values to just under 10 x 10^3 features, considering all separate frequencies.

Up to this point all signal features or markers were decomposed according to sensor, axis and frequency. From this point on all frequencies belonging to a single sensor axis have been grouped together. This was done because it no longer made any kind of sense to make a distinction between signals which were grouped according to their time domain frequency, now being transformed to space domain.

Despite this fact not all extracted features were used. Due to slow convergence speed of the COMB Genetic Signal Shift Adjustment algorithm, only features having the width within the domain [24, 28] m have been considered. It is because of this that the final signal density for the 45-km test line between Cluj-Napoca and Gherla counted less than 2500 features and allowed for reasonable convergence.

5. Signal Shift Adjustment

The problem with the decomposed dyadic signals is that, although they correlate inside the same journey, they do not exhibit the same behavior across different travels, since they account for different collection conditions.

The fact that correlation between train journeys is so small follows directly from train localization accuracy, since inside a single journey a common location is calculated for all devices, with the above-mentioned MAX-AV algorithm [12]. This is not the case of multiple distinct journeys, for which the possibility of inter-journey position synchronization does not exist. As a result, the location of a track feature is a Gaussian distribution centered around its common average.

Before describing the time complexity of our algorithm, it would be judicious to present the limiting factors of using multiple train journeys. Since they are time ordered one would expect a relation of associativity to exist between them, while a lack of correlation between any consecutive journeys would be enough to entail major modifications to the track and require the re-

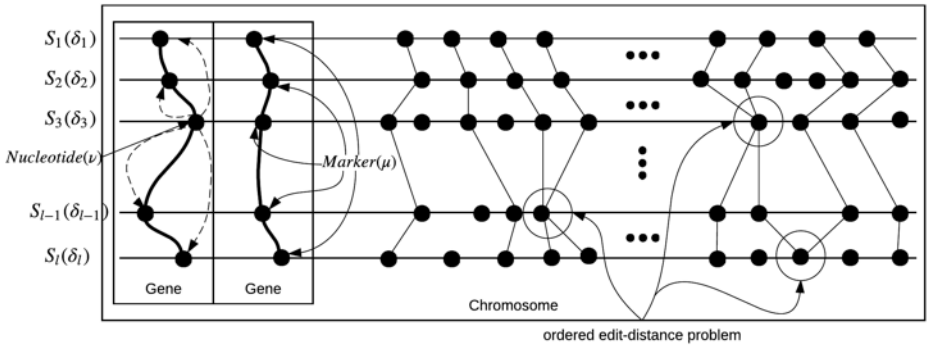

Figure 5. Encoding of the model of signal shift adjustment problem by using a string edit-distance genetic algorithm concepts

computing of the train journey stack $l = \{l_{n-1}...l_1\}$; But this is not the case since a break in correlation could as easily describe a failure in the collection. Thus, the complexity would not linearly increase with the size of the journey stack $|l|$, but require that all combinations C_l^2 be explored to rule out any localized failures.

The size of $|l|$ is considerably smaller than the complexity $O(k * n^2)$ of the tree edit-distance class of algorithms, where n is the size of the compared signals and k is a large multiplicative generic factor, made to hide the in-degree parameters. However, it still has an exponential growth, pertaining to an increase of $|l + 1|$ in execution complexity with each increase in the number of possible combinations C_{l+1}^2 .

As a result, it was decided that it would be best to use a stochastic algorithm to be able to calculate at least a sub-optimal solution to the signal shift adjustment problem. The name selected for this algorithm is COMB, and it provides a sorted alignment of signals, making sure none of the matched Humps are tangled (Figure 5).

The solution to the signal shift adjustment problem was found in designing a genetic based approach to the problem. First, we encoded all the possible edit-distance solutions of string adjustments into genomes. Then, we defining the basic operations of mutation and crossover to fit the problem of signal adjustment. Finally, by using a fitness based result selection approach we could solve the problem of unequal position shifts of the signal.

The unequal signal shift problem, as described here, is not a trivial one since the δ used in Figure 5 is not a trivial variable. It is actually a function $\delta(t)$, which varies with time and has a random nature, as shown in the MAX-AV algorithm [12].

5.1 Encoding Humps into Genomes

The idea of using GA's (Genetic Algorithm) to solve the problem of signal shift adjustment first arose when comparing the relative similarity of the concepts of our problem to the abstractions present in GA's (Figure 5).

The basic signal structure, the hump or marker, represented by a single dot in Figure 5, contains no information with regard to fitness. It consists only of a set of variables (p, w, m), which represent the position, the width and magnitude of the signal. Each of these 3 variables are used by the more elevated structures to calculate fitness. The only other information that a hump contains is knowledge of the string or signal that contains it and its location within the string. These, as seen later, will offer advantages to performing crossover and mutation operations.

The first of these components is the nucleotide, denoted with υ. It represents the same information as the marker but has additional knowledge about matched markers in the sibling journeys (which the chromosome tries to align) and contains information about fitness.

The nucleotide uses the w and m variables, which also have information about the probability distribution of the stochastic data they contain. It is with the help of the probability distribution, that the fitness is calculated. In our case, the fitness of the nucleotide represents the probability that the referenced marker belongs to its Gene or to the shift adjusted neighborhood.

The way in which this is calculated is by using a normalized p-value of a 2-tailed test under the null hypothesis, which states that the value belongs to the above-mentioned Gene distribution. The requirement of using a 2-tailed test stems from the fact that matching must occur from both sides of the distribution bell. The normalization is required because the p-value is a threshold value denoting the amount of cumulative probability that is before or after the value on its axis. The formula for calculating the p-value and the normalized *p'*-value is given by:

$$p(x) = 2 \times \min\left\{\Pr\Pr(X \geq x/H), \Pr(X \leq x/H)\right\}$$

$$\forall x \in \upsilon(p, w, m)$$

$$p'(x) = 1 - p(x)$$

It is noteworthy to mention that the fitness algorithm for nucleotides uses 2 types of distributions to model the p-value. The first is the T-Distribution, which is used when dealing with small amounts of data. The second is the Gaussian Distribution, which is used when the amount of data used to calculate the mean and standard deviation is above 32 values, as described in the Statistical Handbook [21].

Moving up in the hierarchy we find the Gene, which is nothing else but a bundle of related Nucleotides ordered according to a central position. It is here that all the fitness of all the variables of all the nucleotides are merged into a single normalized value in the range of [0, 1]. The values are first merged among the variables of a certain type, i.e. width variables and magnitude variables. Only then when the fitness of all the variables are condensed into 2 values corresponding to width and magnitude, does the final merging occur. Both steps employ Fisher's method for merging the data as described below.

What the method does, is to combine multiple p-values into a χ^2_{2k}, with twice the amount of degrees of freedom as the number of p-values merged by using the following formula:

$$\chi^2_{2k} \sim -2\sum_{i=1}^{k} ln\,(p_i)$$

$$\chi^2_{2k} = \sum \frac{(observed - expected)^2}{observed}$$

Each Gene after calculating their local fitness, adjusts it by multiplying it with the monotonicity factor of the feature. The factor used is a local "distance of monotonicity" calculated by using the following formula:

$$f = 1 - \frac{2}{k} * dist$$

where *k* is the number of elements in the localized window and is the distance of the element from its expected location in the global list. When adjusting with the monotonicity factor, the fitness is de-normalized and falls in the open range of [1 (-*k*/*n*).

Then, finally, all the fitness values of all the genes of a chromosome are summed together to obtain a value which is directly related to the size of the chromosome.

5.2 Mutation and Crossover of Signal Edit Distances

Before delving into the mutation and crossover operations it is important to note that the edit-distance problem class, as highlighted in Figure 5, is an ordered problem, but the algorithm tries to solve for an unordered state. This is helpful for gaining genetic variance and designing more general crossover operations.

The mutation and crossover operations are the most important operations of GA. They allow the algorithm to converge to a nearly-optimal solution. But there are certain pitfalls to keep in view. The first is a slow convergence speed of the algorithm, which happens when the algorithm keeps searching even when it clearly has good solutions at its disposal. This can happen if a too large mutation and crossover rate are selected. On the other hand, there is the problem of a too fast convergence, where the algorithm loses interest in searching, due to its small mutation and crossover rates.

No self-optimizing approach was used to establish the optimal values for mutation and crossover rates due to the inherent time complexity involved in calculating a single iteration of the approach. Instead an empiric trial and error technique was used to establish the optimal values.

Before starting to apply the evolutionary operators on the population, one must make sure that the population is sufficiently heterogeneous, from a genetic standpoint. Because just as in nature, one can significantly hamper evolution in our case, search if the starting population is too small or not genetically diverse enough. As a result, having a large enough population is a good starting point.

Mutation is a simple operation designed to offer an algorithm reset, to eject the algorithm out of a local minimum and allow it to continue exploring the state space. The solution envisioned in this work uses a single operation randomly shifting the Nucleotides composing the Gene left or right, with at most 3 positions.

The crossover operation is designed to combine two Chromosomes together by merging each corresponding Gene, based on their index. The Genes are combined by calculating 2 possible monotonicity improvements of each of their corresponding Nucleotides, by using linear interpolation and extrapolation. Once the directions are established shifting the best Nucleotide is shifted in that specific direction with one position.

The crossover also treats the case in which the Nucleotide has a high enough monotonicity and fitness to warrant any further changes, as a probability of change based on the inverse fitness of the Nucleotide adjusted with monotonicity. For values smaller than 0, there is a 100% probability of the crossover operation to be performed.

This approach is at the heart of the evolutionary algorithm since it allows for the GA to search the opposing Gene's neighborhood for better solutions.

5.3 ZeroMarker as a Placeholder for Missing Data

There are cases which have been foreseen, where Markers or Nucleotides of one journey cover a bigger chunk of track than their counterparts, or some features are not detected in all journeys; there should be the possibility of marking missing data. For this case, a special Nucleotide has been devised, named the ZeroMarker. Its only mission is to mark missing data, while at the same time not losing information about position and possible neighboring Humps.

Both crossover and mutation operations define the special case of ZeroMarkers, by propagating or generating the value as a counterbalance to poor fitness values, so that Nucleotides with a fitness smaller than -5 are replaced.

ZeroMarkers have a fitness equal to 0 and do not contribute to the overall fitness of the Chromosome. In case of poor correlation, it can happen that most of the Nucleotide fitness values are smaller than 0, so that a ZeroMarker would be a positive gain to the algorithm.

To avoid such cases where all Nucleotides would be converted to ZeroMarkers, a propagation function has been defined for both crossover and mutation operations, limiting the shelf life of the HumpMarker.

5.4. Result and Selection Approaches

The results of running the evolutionary algorithm were obtained by using a tournament selection over an elitist population. What this means is that a number x of elements were selected from the pool of chromosomes and run through a tournament, where only the fittest was selected. By using this technique, a reasonable chance was given to sub-optimal solutions, with lower fitness, to reach the next round since the tournament size was empirically selected to include only 10 competitors per round, given a population of just 100 chromosomes.

The elitist population assured that the algorithm would only select the fittest chromosome in the tournament selection, which would cause the algorithm to converge.

The results of this evolutionary approach using GA to align the signals across multiple journeys allowed for a high correlation to occur and thus allow for a signal averaging approach to collapse the signals into a single feature vector to be used in recognition.

6. Conclusions and Future Work

This work has presented thus far the problem of modeling acceleration and gyroscopic signal patterns connected to track line representation. The basic idea was to find a representation which

was both easy to process and small enough to transfer over low speed networks in a matter of a few minutes. Additionally, the problem of completeness and correctness of signal representation was also considered, but only within the context of similarity of the decomposed signal to the Daubechies 4-tap wavelet, which was used to extract the dyadic frequency from the original signal.

Finally, the work concluded with the COMB algorithm, a signal shift adjustment algorithm, designed to reduce the inconsistencies produced by the GPS positioning system in transforming the signal form the time domain to the one dimensional along-track space domain.

The stochastic nature of the algorithm allowed it to obtain a sub-optimal result in reasonable time, since a deterministic approach neared $O(n^3)$ execution complexity, for large sets of train journeys. Not only that but it permitted the use of fuzzy comparison approaches, in which two features exhibited similarity within the range of [0 - 1]. Thus offering a natural comparison of the inherently continuous (p, w, m) variables.

The results obtained allowed for the signal to correlate inside and across journeys. This demonstrates that the algorithm represents a viable solution to the problem of signal recall across multiple train journeys and allows for a database of features of the track line to be formed. This data structure could then be used in an online localization solution based on a particle filter approach, such that the features identified in the offline processing could then be recognized and thus offer a notable improvement to the localization algorithm.

The improvement would come from the reduced reliance on GNSS solutions but also from heightened accuracy provided by the generations of data collected during the offline runs of the above presented GA approach.

Acknowledgment

This paper is supported through the Sectoral Operational Programme Human Resources Development (SOP HRD), ID/134378 financed from the European Social Fund and by the Romanian Government.

The research has been performed at the Technical University of Cluj-Napoca and was funded by the Rail Automation business unit of Siemens AG. The views expressed are those of the authors and not necessarily those of Siemens AG.

REFERENCES

1. Acharya, A., Sadhu, S. & Ghoshal, T. (2011). Train localization and parting detection using data fusion, *Transportation Research Part C: Emerging Technologies*, *19*(1), 75–84.

2. Astarita, V., Caruso, M. V., Danieli, G., Festa, D. C., Giofre, V. P., Iuele, T. & Vaiana, R. (2012). A mobile application for road surface quality control: Uniqualroad. In *Procedia - Social and Behavioral Sciences*, *proceedings of EWGT2012 - 15th Meeting of the EURO Working Group on Transportation*, *54* (pp. 1135 – 1144).

3. Aucouturier, J. J. & Sandler, M. (2002). Finding repeating patterns in acoustic musical signals: Applications for audio thumbnailing. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society.

4. Barbu, G., Hanis, G., Kaiser, F. & Stadlmann, B. (2014). SATLOC-GNSS based train protection for low traffic lines, *Rail Signalling and Telecommunication*.

5. Belabbas, B., Grosch, A., Heirich, O., Lehner, A. & Strang, T. (2013). Curvature classification for trains using along-track and cross-track accelerometer and a heading rate gyroscope. In *Proceedings of the European Navigation Conference*.

6. Bille, P. (2005). A survey on tree edit distance and related problems, *Theoretical Computer Science*, *337*(1), 217 – 239.

7. Brooks, C. A. & Iagnemma, K. (2005). Vibration-based terrain classification

for planetary exploration rovers, *IEEE Transactions on Robotics*, *21*(6), 1185–1191.

8. Broquetas, A., Comeron, A., Gelonch, A., Fuertes, J. M., Castro, J. A., Felip, D., Lopez, M. A. & Pulido, J. A. (2012). Track detection in railway sidings based on MEMS gyroscope sensors, *Sensors*, *12*(12), 16 228–16 249.

9. Bychkovsky, V., Chen, K., Goraczko, M., Hu, H., Hull, B., Miu, A., Shih, E., Zhang, Y., Balakrishnan, H. & Madden, S. (2006). The cartel mobile sensor computing system, *SenSys*, *6*, 383–384.

10. Colceriu, V. D., Bacu, V., Stefanut, T. & Gorgan, D. (2014). Distributed Context Aware Train Localization System. In *International Conference on Intelligent Communication and Processing* (pp. 417-422).

11. Colceriu, V. D., Bacu, V., Stefanut, T. & Gorgan, D. (2015). Generating accuracy and integrity aware train movement maps using GNSS and MEMS sensors. In *International Conference on Intelligent Communication and Processing* (pp. 441-446).

12. Colceriu, V. D., Bacu, V., Stefanut, T. & Gorgan, D. (2017). Low grade sensor data based annotation of topological railway maps. In Dumitrache I., Florea, A. M., Pop F. & Dumitrascu, A. (Eds), *21st International Conference on Control Systems and Computer Science* (pp. 167–174).

13. Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets, *Communications on pure and applied mathematics*, *41*(7), 909–996.

14. Diggelen, F. van (2007). GNSS Accuracy: Lies, Damn Lies, and Statistics, *GPS World*.

15. European Railway Traffic Managment System /European Train Control System, *Baseline 3*. (2011). European Railway Agency Std.

16. *Federal Railroad Administration*. (2017). "PosiPTC Overview & Individual Railroads", Association of American Railroads. [Online]. Available: <https://www.fra.dot.gov/ptc>.

17. Heirich, O., Lehner, A., Robertson, P. & Strang, T. (2011). Measurement and analysis of train motion and railway track characteristics with inertial sensors. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (pp. 1995–2000).

18. Heirich, O., Steingass, A., Lehner, A. & Strang, T. (2013). Velocity and location information from onboard vibration measurements of rail vehicles. In *2013 16th International Conference on Information Fusion (FUSION)* (pp. 1835–1840). IEEE.

19. Iliopoulos, C. S. & Kurokawa, M. (2002). String matching with gaps for musical melodic recognition. In *Proceedings of the Prague Stringology Conference* (pp. 55–64).

20. Johnsson, R. & Odelius, J. (2012). Methods for road texture estimation using vehicle measurements. In *Proceedings ISMA 2012: International Conference on Noise and Vibration Engineering: including USD2012:Leuven*, *17 - 19 September 2010* (pp. 1573–1582). Katholieke Universitat.

21. *NIST/SEMATECH* (2012). Engineering statistics. [Online]. Available: <http://www.itl.nist.gov/div898/handbook/index.htm>.

22. Pachl, J. (2013). Fendrich, L. & Fengler, W. (Eds.), Betriebsfuhrung der Infrastruktur, *Handbuch Eisenbahn-infrastruktur*, 405–440. Springer Berlin Heidelberg.

23. Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D. & Srivastava, M. (2010). Biketastic: sensing and mapping for better biking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1817–1820). ACM.

24. Scholten, H. & Bakker, P. (2011). Opportunistic Sensing in Train Safety Systems, *International Journal on*

Advances in Networks and Services, *4*(3-4), 353–362.

25. Stadlmann, B., Kaiser, F. & Maihofer, S. (2012). Rechnergestutztes Zugleitsystem fur die Pinzgauer Lokalbahn, *Signal+Draht*, *104*, 28– 33.

26. Stadlmann, B., Mairhofer, S. & Hanis, G. (2010). Field experience with GPS based train control system. In *Proceedings of GNSS 2010-The European Navigation Conference on Global Navigation Satellite Systems*.

27. Tiberius, C. & Verbree, E. (2004). GNSS positioning accuracy and availability within Location Based Services: The advantages of combined GPS- Galileo positioning. In Granados G. S. (Ed.), *2nd ESA/Estec workshop on Satellite Navigation User Equipment Technologies* (pp. 1–12). ESA publications division, Noordwijk.

28. Zimmerman, J., Tomasic, A., Garrod, C., Yoo, D., Hiruncharoenvate, C., Aziz, R., Thiruvengadam, N. R., Huang, Y. & Steinfeld, A. (2011). Field Trial of Tiramisu: Crowd-sourcing Bus Arrival Times to Spur Co-design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *ser. CHI '11, New York, NY, USA: ACM, 2011* (pp. 1677–1686).