international workshop on

Maritime Anomaly Detection

MAD 2011 Workshop Proceedings

June 17, 2011, Tilburg, The Netherlands

http://mad.uvt.nl

Contents

Maritime Anomalies in Tilburg

Welcome to the first international workshop on Maritime Anomaly Detection (MAD 2011). With the increasing availability of sensors and advanced computational resources, the semi-automatic detection of anomalies in digital data has become a prominent challenge. The MAD workshop focuses on the maritime domain, in which large volumes of data on vessels need to be monitored to prevent accidents or terrorist attacks from happening. Digital detection methods aim at supporting operators in their detection and identification of anomalies in maritime data by relying on combinations of maritime data sources (e.g., AIS data, radar data, and web-based data).

The main challenge of maritime anomaly detection is to find proper combinations of maritime domain knowledge for properly representing maritime data and effective outlier detection algorithms. The contributions to the MAD workshop report on such proper combinations that emphasize (1) visualization, (2) vessel-behavior representation, and (3) semi-automatic detection. Visualization of properly pre-processed data support operators in their visual detection of anomalous patterns. Effective ways of representing the behavior of vessels allow for the definition of appropriate metrics to identify anomalous vessel behaviors. Density-based outlier detection algorithms employing these metrics on the behavioral vessel data enable the semi-automatic identification of suspicious behaviors.

The workshop features two keynote speakers: David Aha of the Adaptive Systems Section of the Naval Research Laboratory in the United States and Maurice Glandrup of the Above Water Systems department of Thales Systems in The Netherlands. Their presentations will provide intriguing complementary perspectives on the scientific development and application of maritime anomaly detection methods. Apart from the keynote presentations, nine oral presentations by international established researchers in anomaly detection offer a broad perspective on the state-of-the-art in the domain. During the poster session, seven contributions offer an additional view on the international research activities.

We hope that the workshop will inspire you to develop and test new methods in the maritime domain.

Jeroen Janssens, Eric Postma, and Joke Hellemons,
Tilburg center for Cognition and Communication
Tilburg University

Workshop Program

Start	Duration	Title	Authors
9:00	0:30	Registration	
9:30	0:15	Opening	Jaap van den Herik, Tilburg center for Cognition and Communication, The Netherlands
9:45	0:45	Keynote: Towards Maritime Behavior Recognition and Anomaly Detection	David Aha, Naval Research Laboratory, USA
10:30	0:30	Density Based, Visual Anomaly Detection	Roeland Scheepens, Niels Willems, Huub van de Wetering, and Jarke van Wijk, Eindhoven University of Technology, The Netherlands
11:00	0:30	ConTraffic: Maritime container traffic anomaly detection	Aristide Varfis, Evangelos Kotsakis, Aris Tsois, Maxym Sjachyn, Alberto Donati, Elena Camossi, Paola Villa, Tatyana Dimitrova, and Muriel Pellissier, European Commission, Joint Research Centre, Italy
11:30	0:30	Comparing Vessel Trajectories using Geographical Domain Knowledge and Alignments	Gerben de Vries, Universiteit van Amsterdam, The Netherlands; Willem Robert van Hage, Vrije Universiteit Amsterdam, The Netherlands; and Maarten van Someren, Universiteit van Amsterdam, The Netherlands
12:00	1:00	Lunch	
13:00	0:30	Keynote: Towards improving situation awareness for operators in the maritime domain	Maurice Glandrup, Thales Naval, The Netherlands
13:30	0:30	Spatio-Temporal Visualisation of Outliers	Laurent Etienne and Cyril Ray, Naval Academy Research Institute, France; and Gavin Mcardle, National Universtiy of Ireland Maynooth, Ireland
14:00	0:30	Maritime Anomaly Detection using Stochastic Outlier Selection	Jeroen Janssens, Eric Postma, and Jaap van den Herik, Tilburg University, The Netherlands
14:30	0:30	Maritime Route Anomaly Detection	Richard Lane, QinetiQ, United Kingdom
15:00	0:30	Break	
15:30	0:30	Applying V-Analytics to AIS Data	Gennady Andrienko and Natalia Andrienko, Fraunhofer Institute IAIS, Germany
16:00	0:30	An Evaluation of Fractal/Velocity Pattern Extraction	René Enguehard, Rodolphe Devillers, and Orland Hoeber, Memorial University of Newfoundland, Canada
16:30	0:30	Incremental Stream Clustering for Anomaly Detection in Maritime Surveillance	Anders Holst and Jan Ekman, Swedish Institute of Computer Science, Sweden
17:00	0:10	Poster pitches	
17:10	1:50	Poster session and drinks	
19:00	?	Workshop banquet	

Towards Maritime Behavior Recognition and Anomaly Detection

David W. Aha
Navy Center for Applied Research in Artificial Intelligence,
Naval Research Laboratory (Code 5514),
Washington, DC 20375, USA,
david.aha@nrl.navy.mil

Our group has been studying methods for automatically recognizing maritime threats from land- and ship-based cameras and other sensors. This is a particularly important task for protecting Navy and commercial personnel and assets. Several challenges exist, including those pertaining to vision processing (which is not our focus), behavior recognition, and threat assessment. I will describe our initial progress on these tasks, including the perspective we use to approach these tasks, a comparison of anomaly detection models for recognizing unusual maritime behaviors, and a comparison of probabilistic relational models for behavior recognition.

Density Based, Visual Anomaly Detection

Roeland Scheepens, Niels Willems, Huub van de Wetering, Jarke J. van Wijk
{R.J.Scheepens, C.M.E.Willems, H.v.d.Wetering, J.J.v.Wijk}@tue.nl
Eindhoven University of Technology, The Netherlands

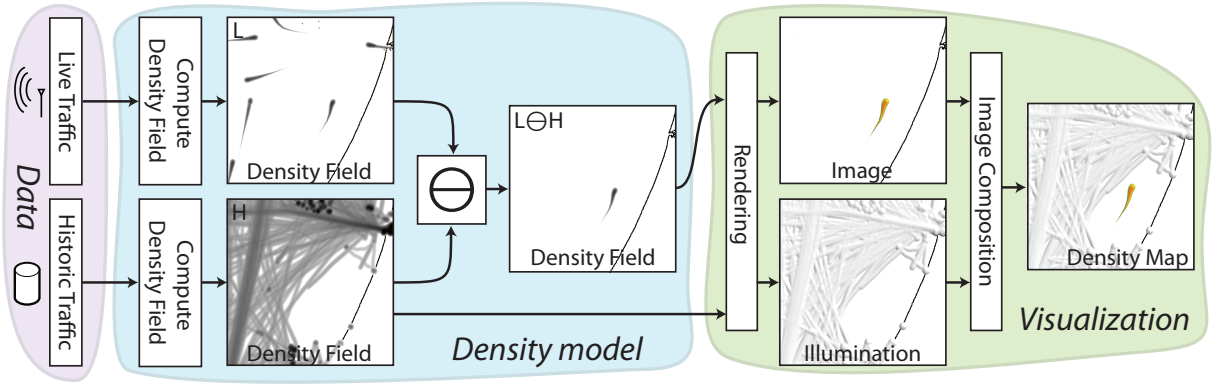


Figure 1: A density field is created for both live (L) and historical (H) data, serving as input for anomaly operator $\ominus$ and resulting in a new density field $L \ominus H$. This field is visualized using a multi-hue color map and composed, with the historical density field, into an illuminated height field.

1 Introduction

Anomaly detection in a visual setting allows a user to more easily recognize and interactively investigate anomalies. We show how density maps by Scheepens *et al.*[1] can support the analysis. Density maps are created in a four-way procedure–see Figure 1: First, the user divides the data into one or more subsets using interactive controls. Second, for each subset, smoothed trajectories are aggregated in a density field. Third, the density fields are combined into a single field using an aggregation operator and, finally, the resulting density field is rendered and composed into a density map. With modern graphics hardware, this process can be performed at interactive speeds.

1.1 Density Model

The trajectory $\boldsymbol{\alpha}$ of an object, defined over a continuous range of time $[t_0, t_1]$, is given at time t as a tuple $\boldsymbol{\alpha}(t)$ containing attributes such as position $\mathbf{p}(t)$ and velocity $v(t)$. A density field is generated by moving a smoothing kernel along the trajectories. A user can create multiple density fields, one for each subset of data defined by an attribute filter $\mathscr{F}_{\boldsymbol{\alpha}}$. Space is subdivided into a regular grid of cells with equal area, for which a density field is computed per cell Q. The density fields are then combined using a per-cell density aggregation and visualized using a multi-hue color map. The combined density field is rendered as an illuminated height field and composed with the color mapped density field into a density map.

Large vessels usually follow predefined shipping lanes. If such a vessel moves outside these shipping lanes, this constitutes a possible anomaly. We can find and investigate such anomalies by comparing a density field H of a sufficiently large set of historical movements with a density field L of live data containing recent movements. Where the historical data represents normal behavior, we define anomalous movements as current movements that differ significantly from historical movement patterns. We introduce an anomaly operator $\ominus$ as a density aggregation and define the density field $(L \ominus H)(Q) = \max(0, \omega L(Q) - H(Q))$ to reveal anomalous areas, where ω is some weight factor.

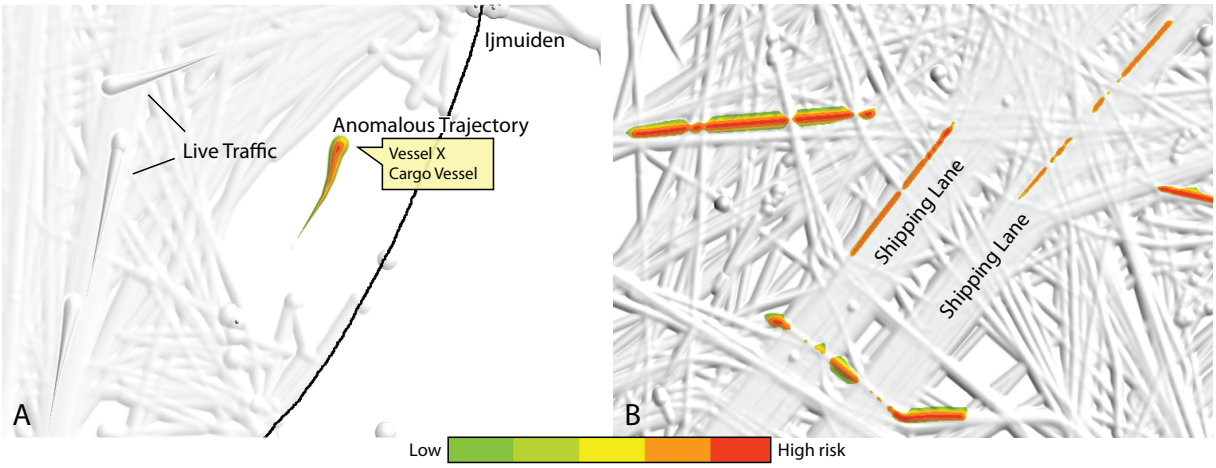


Figure 2: Vessel traffic in front of the Dutch coast. (A) Live traffic with a half hour tail. The colored trajectory is detected as an anomaly in the context of historical data, which turns out to be a single cargo vessel after inspection. (B) Anomalies among vessels carrying potentially hazardous material in color and the context of the entire historical set in gray.

2 Use Case: Anomaly Detection

We investigate a small area North of Rotterdam. For our live data set we use half an hour of trajectories, for our historical data set we use six days of trajectories and we set the weight ω to 1. The smoothing kernel varies over time such that older movements are displayed smaller, indicating the direction. Anomalies are visualized using a green-to-red color map and a combination of the historical density field H and live density field L is shown as context using illumination–see Figure 2A. We now see an anomalous area worth investigating by looking at the colors. We can click on this area to retrieve a list of vessels contributing to this anomaly. This tells us that the anomaly is caused by a cargo vessel. This vessel is likely outside of the designated shipping lanes to resupply ocean platforms in the area.

In Figure 2B we assign a filter $\mathscr{F}_{\boldsymbol{\alpha}}$ such that only vessels carrying potentially hazardous material are considered. A density field of the entire historical data set is shown as context. We see several vessels moving in areas, mostly outside the shipping lanes, where there are normally no or only few vessels carrying potentially hazardous material.

3 Future Work

We have concentrated on spatial anomalies, however, in future research we hope to define more operators for other types of anomalies such as drifters, speeders, or vessels moving against the main traffic flow. Furthermore, more complicated anomalies such as behavioral patterns within single trajectories or interactions between vessels may be recognized and visualized as anomalies.

This work has been carried out as a part of the Poseidon project at Thales Nederland under the responsibilities of the Embedded Systems Institute (ESI). This project is partially supported by the Dutch Ministry of Economic Affairs under the BSIK program.

[1] Roeland Scheepens, Niels Willems, Huub van de Wetering, and Jarke J. van Wijk. Interactive visualization of multivariate trajectory data with density maps. In *Proceedings of IEEE Pacific Visualization Symposium*, pages 147–154, 2011.

ConTraffic:Maritime container traffic anomaly detection

A. Varfis, E. Kotsakis, A. Tsois, A.V. Donati, M. Sjachyn, E. Camossi, P. Villa, T. Dimitrova, M. Pellissier
European Commission, Joint Research Centre, Ispra, Italy
{aristide.varfis, evangelos.kotsakis, aris.tsois, maxym.sjachyn, elena.camossi, paola.villa, tatyana.dimitrova, muriel.pellissier}@jrc.ec.europa.eu
alberto.donati@ext.jrc.ec.europa.eu

1 Introduction

Although most of the world's cargo is transported by means of maritime containers, only a small percentage of this traffic is physically inspected [1]. This, inevitably, leads to the possibility of illegal activities such as evasion of customs duties, weapon and drug smuggling, etc. ConTraffic is a research prototype constructed to probe the idea of illicit cargo detection via trip related information analysis and assist authorities in detection of such cargo. The primary operational hypothesis states that: actions related to the transportation of illicit cargo often disturb a normal trip pattern and could thus cause anomalies. The system has been successfully used within the framework of mutual assistance between customs to identify false declarations of origin and smuggling of goods.

2 The ConTraffic dataset

In order to be able to perform any kind of anomaly detection the ConTraffic system has to first build its dataset. ConTraffic collects its data on container trips from heterogeneous, publically available sources. The collection is done through web-spiders which extract the information from the deep-web and store it in the data staging area of the ConTraffic data warehouse. The integration of the collected data into a coherent dataset requires significant semi-automatic transformations and cleaning that deals mostly with non-standard text strings defining geographic locations and container event types.

The resulting dataset is stored in a data warehouse which facilitates the analysis processes by using appropriate data structures and indexes. The dataset contains information in the form of container events. Each event defines what happened to a particular container on a particular date at a particular location. Many events also mention the vessel name involved. Currently the dataset contains more than 900 million events about more than 12 million containers.

3 Anomaly detection

Defining what constitutes an anomalous container trip is not a trivial task. In this section we present some of the typologies of anomalies which we have examined in ConTraffic while in the next section we briefly mention our on-going work on new techniques.

Carrier companies that transport containers worldwide do not connect directly every possible pair of origin and destination ports. Therefore, many containers need to be transshipped during their maritime trip. As transshipments have a non-negligible cost, one would expect a carrier company to avoid superfluous transshipments. Our dataset enables us to reconstruct, from the large dataset of container events, the trips of vessels and scan for instances of superfluous transshipments. A transshipment is considered superfluous when the trip of the container could have been executed without it. Although logistical reasons could explain some of the superfluous transshipments, one rare form, which we call a *loop* is particularly suspicious. It occurs when the second vessel of the trip, which loads the container under scrutiny at the transshipment port, subsequently comes back to the port of origin along its route to

the final destination port. The detection of such anomalies has been implemented by algorithms in MATLAB. The algorithms reconstruct container and vessel trips from the dataset of container events dealing with the issues of incomplete and sometimes inconsistent information. These types of anomalies are also studied in the context of the Maritime Container Ontology (MCO) [2]. The anomaly patterns have been formalized as logical axioms in the MCO, which are used to check the consistency of the container and vessel itineraries and to identify anomalous consignments.

A different approach is the Timing Factors Anomaly Analysis (TFAA). The purpose of TFAA is to select the container trips manifesting an anomaly with respect to some trip factors. The factors considered are the following: a) number of transshipments; b) the total transshipment time; c) the time from *gate-in* to *load on board* (i.e., the time elapsed between the moment the container enters the port and the moment it is loaded on the vessel); and d) the time from *discharge* to *gate-out* at final destination (i.e., the time elapsed between the moment the container is unloaded from the vessel and the moment it leaves the port). The anomaly score of each of the above mentioned timing factors is defined as follows:

$$AS(x) = \frac{\exp^{-p(x)} - \exp^{-1}}{1 - \exp^{-1}}$$

where $p(x)$ is the relative frequency of the observed value x, and as $AS(x)$, assumes values in $[0,1]$. It is worth noticing that this definition is totally independent of any assumption about the distribution type, which in fact might have any arbitrary shape.

4 On-going research

It appears that graphs are well adapted to represent container trip information and several algorithms exist to detect anomalies in graphs [3, 4]. In this direction we are investigating how such techniques could be used to identify containers with high probability of carrying illicit cargo. We are investigating also how geographical representation and visualization techniques can assist in detecting anomalies. We believe that an appropriate geographical visualization of the multi-dimensional ConTraffic dataset could provide highly valuable assistance for data analysis [5], data understanding and detection of anomalies. In this direction we are studying how to best display the information in different overlays by using Google Maps and Google Earth, allowing a high degree of interaction to the user.

There are also many other directions in which we intend to extend our research. For example: What is an efficient algorithm to detect superfluous transshipments in large datasets like ours? How kernel-based anomaly detection approaches can be extended to spatio-temporal data processing, in order to better reveal suspicious container trips?

References

[1] R. Hoshino, et al. *Two-stage Approach for Unbalanced Classification with Time-varying Decision Boundary: Application to Marine Container Inspection*, ISI-KDD 2010, ACM.

[2] P. Villa and E. Camossi. *A description logic approach to discover suspicious itineraries from maritime container trajectories*. GEOS 2011, Vol. 6631 of LNCS, p. 182-199. Springer-Verlag.

[3] W. Eberle and L. Holder. *Detecting Anomalies in Cargo Shipments Using Graph Properties*. ISI 2006, IEEE.

[4] C. Noble and D. Cook. *Graph-Based Anomaly Detection*. SIGKDD ICKDM 2003, p. 631-636, ACM.

[5] T. Melanie and T. Moller, *Human factors in Visualization Research*, IEEE Trans. on Visualization and Computer Graphics, 10(1), 2004.

Comparing Vessel Trajectories using Geographical Domain Knowledge and Alignments

Gerben K.D. de Vries[†] Willem Robert van Hage[*]
Maarten van Someren[†]
[†]Informatics Institute, University of Amsterdam, Sciencepark 904, 1098 XH,
Amsterdam, the Netherlands, {G.K.D.deVries|M.W.vanSomeren}@uva.nl
[*]Computer Science, VU University Amsterdam, de Boelelaan 1081a, 1081 HV,
Amsterdam, the Netherlands, W.R.van.Hage@vu.nl

In this paper [1] we present an alignment based similarity measure that combines low-level vessel trajectories with geographical domain knowledge, such as the name and type of the regions that vessels pass through and stop. We use this similarity measure in a clustering experiment to discover interesting behavior and in a classification task to predict the type of the vessel for a trajectory. The combination of information gives the best average classification accuracy. For both clustering and classification we use kernel based algorithms.

We define a trajectory as $T = \langle x_1, y_1 \rangle, \ldots, \langle x_n, y_n \rangle$, ignoring the temporal dimension. The number of vectors in T is denoted as: $|T|$. In the *stop* and *move* model of [5], the trajectories in our experiment are moves. They are delimited by the vessel entering the area of observation or starting, and the vessel leaving the area of observation or stopping.

The geographical domain knowledge comes as two simple ontologies. One, **A&C**, contains the definitions of different anchorages, clear ways, and other areas at sea. The other ontology, **H**, defines different types of harbors, such as liquid bulk and general cargo. For both ontologies, we created a SWI-Prolog webservice [6] to enrich vessel trajectories with geographical features. The first service returns a set of specific type, label pairs corresponding to the regions in **A&C** that intersect with a given point. We create a sequence of sets of geo-labels $T^L = L_1, \ldots, L_{|T|}$ for a trajectory T with this service. For the start and end of a trajectory we define objects that contain information whether the vessel is stopped and if so at what harbor or region. We discover this harbor using the second webservice, which matches a point to the nearest harbor in **H** that is within range and returns the label and specific type of this harbor. If there is no harbor close, we use the first webservice.

For the sequences T and T^L we compute similarity using an edit distance alignment, which we discovered in previous work [2] to perform the best on a vessel trajectory clustering task. To compute an edit distance, we need a substitution function and a gap penalty. The substitution function for trajectories T is defined as: $\mathbf{sub}_{\text{traj}}(\langle x_i, y_i \rangle, \langle x_j, y_j \rangle) = -\|\langle x_i - x_j, y_i - y_j \rangle\|$, i.e. the negative of the Euclidean distance. We take the value for the gap penalty g from the mentioned previous work. For T^L, the substitution function $\mathbf{sub}_{\text{lab}}(L_i, L_j)$ expresses how many labels the sets of labels L_i and L_j have in common. We set g as the minimally possible $\mathbf{sub}_{\text{lab}}$ score.

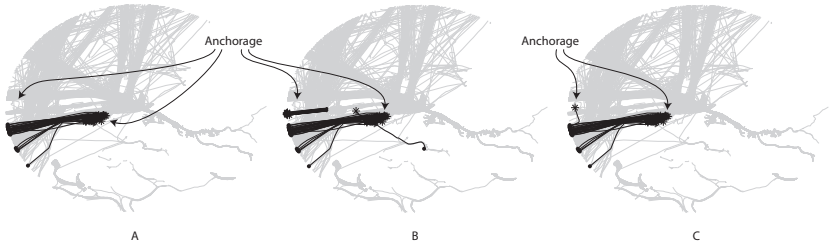


Figure 1: Example of a cluster of trajectories showing anchoring behavior. The example cluster is shown in black against the entire dataset in gray. The start of trajectory is indicated by a dot, the end by an asteriks.

The similarity $Sim(S,T)$ between two sequences S and T is the score of the alignment that has the maximum edit distance score for all possible alignments between these sequences, divided by $|S|+|T|$ to give the average score per element. In the experiments we use kernel based algorithms. For all sequences T_i and T_j in a set of sequences $\mathscr{T}$, we compute a kernel K as: $K(i,j) = Sim(T_i,T_j)$, then we normalize K and turn it into a kernel by $K = 1 - \frac{K}{\min(K)}$. For trajectories T we get a kernel K_{traj} and for sequences of sets of geo-labels T^L we get a kernel K_{lab}. The similarity between two start/end objects can immediately be put into kernel form and is determined by whether the vessel is stopped or not and how much labels there are in common. This gives us a kernel K_{start} for the start objects and a kernel K_{end} for the end objects. $K_{\text{all}} = w_1 K_{\text{traj}} + w_2 K_{\text{lab}} + w_3 K_{\text{start}} + w_4 K_{\text{end}}$ combines the four kernels above. Clearly, this kernel is symmetric, but it is not guaranteed to be positive semi-definite.

Our experimental dataset consists of 1917 vessel trajectories in a 50km radius area around the Port of Rotterdam, collected using the Automatic Identification System (AIS). The trajectories are compressed with the algorithm in [4], reducing the data by 95%, thus reducing computation time drastically. This compression improves performance on a vessel trajectory clustering task [2] using the same alignment.

For the clustering experiment we used weighted kernel k-means [3], with $k = 40$. We created kernels for 3 different weight settings of K_{all}: equal combination of domain knowledge and raw trajectories, K_{comb}, only raw trajectory information, K_{raw}, and only domain knowledge, K_{dom}. This results in a number of interesting clusters. In Figure 1A we see a cluster from clustering with K_{comb} that shows trajectories that enter the area from the west and anchor in one specific anchoring area. In B and C we plotted the most similar cluster from clustering with K_{raw} and K_{dom}, respectively. In Figure 1B there are also trajectories included that do not show the anchoring behavior, because we only consider raw trajectory information. We see the opposite in Figure 1C, where we have only anchoring behavior, but in different anchoring areas.

We also did a classification experiment, predicting the vessel's type. In total there are 18 types, available from AIS. For classification we used a support vector machine (SVM), with the same kernels as for clustering, in a 10-fold cross validation set-up. The classification accuracy for K_{all} was 75.4%, for K_{raw} 72.2%, and for K_{dom} 66.1%. All results differed significantly under a paired t-test with $p < 0.05$.

The similarity measure that we defined was applied in a clustering task and we gave an example of discovered interesting vessel behavior that is a combination of both raw trajectories and geographical information. We also used the measure in classification to predict vessel types where the combined similarity showed the best performance in terms of classification accuracy. We plan to apply the measure in the task of outlier detection to discover strange vessel behavior.

References

[1] G. de Vries, W. R. van Hage, and M. van Someren. Comparing vessel trajectories using geographical domain knowledge and alignments. In W. Fan, W. Hsu, G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, editors, *ICDM Workshops*, pages 209–216. IEEE Computer Society, 2010.

[2] G. de Vries and M. van Someren. Clustering vessel trajectories with alignment kernels under trajectory compression. In J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, editors, *ECML/PKDD (1)*, volume 6321 of *Lecture Notes in Computer Science*, pages 296–311. Springer, 2010.

[3] I. S. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors – a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1944–1957, 2007.

[4] J. Gudmundsson, J. Katajainen, D. Merrick, C. Ong, and T. Wolle. Compressing spatio-temporal trajectories. *Computational geometry*, 42(9):825–841, 2009.

[5] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. F. de Macêdo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146, 2008.

[6] W. R. van Hage, J. Wielemaker, and G. Schreiber. The space package: Tight integration between space and semantics. *T. GIS*, 14(2):131–146, 2010.

Towards Improving Situation Awareness for Operators in the Maritime Domain

Maurice Glandrup
Thales Naval Nederland B.V.,
P.O. Box 42, 7550 GD Hengelo, The Netherlands
maurice.glandrup@nl.thalesgroup.com

Operators in the maritime domain have to cope with increasingly complex situations at sea to maintain government determined levels for safety and security. A closer analysis of vessel movement before the Dutch coast shows that a considerable part of the vessels do not follow the sealanes in the area. Because of the amount of vessel movement, it is difficult to tell which vessels adhere to traffic rules and which ones do not. In the past years this problem has become more appearent, because more vessels have equipment to broadcast their last position or are detected by sensors. In addition to that, besides spatial-temporal data of vessels, information about the crew, cargo and owner of vessels has become more important. The amount of information can easily result in an overload which causes missing or too late noticing of events that can evolve into kinematical anomalies, information anomalies and their combination. In this talk, some results and the application of techniques that were developed in the research project Poseidon to improve the situation awareness for operators are presented. The techniques range from advanced visualization concepts to semantic reasoning to (semi-) automated detection of anomalies. Highlighting anomalies in a domain as complex as the maritime domain is, however, difficult. To increase the success rate the addition of domain knowledge is helpful. Within Poseidon, domain knowledge is used to help in the identification of anomalies. In the future, we hope to further improve the anomaly detection capabilities for maritime systems. As conclusion of the talk a case is illustrated of a type of anomaly that at this moment is difficult to detect but that should be detectable in the near future.

Spatio-Temporal Visualisation of Outliers

Laurent ETIENNE, Cyril RAY
Naval Academy Research Institute
Lanveoc-poulmic, France
firstname.lastname@ecole-navale.fr

Gavin McARDLE
National Centre for Geocomputation
National University of Ireland Maynooth, Ireland
gavin.mcardle@nuim.ie

1 Data Mining and Pattern Recognition in Movement Data

The increase of maritime location-based systems broadcasting information about ship movements is providing large sets of positioning data. Detecting outliers that behave in an unusual way in such large amounts of data is an active research field linked to data mining, statistical analysis and geovisual analytics. Assuming that moving objects following the same itinerary behave in a similar way, these behaviours can be derived by data mining on spatio-temporal databases (*STDB*). This allows for the understanding of common trajectories (considered as the normality) over a long period of time from which location dissimilarities should raise attention. Such trajectory analysis tied with a geovisualisation process is essential for safety applications. This allows traffic operators to focus on possible outliers and reduce cognitive load in overcrowded areas. The following details a process to qualify the position and trajectory of a moving object both on spatial and temporal criteria and discusses the benefits of geovisualisation for pattern understanding. Figure 1 presents the functional process used to extract spatio-temporal patterns

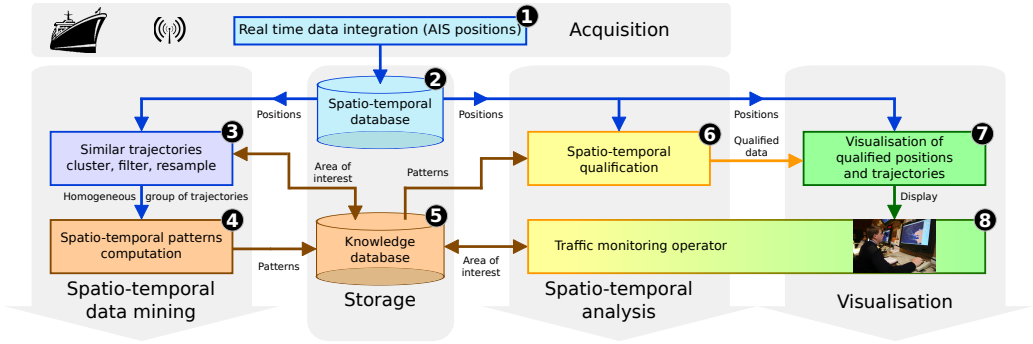


Figure 1: Functional process

from the *STDB* and qualify ship positions and trajectories [3]. A data integration step (Fig. 1, step 1) relies on a monitoring system providing real-time AIS data integration [2]. Trajectories derived from AIS data are clustered in order to get a Homogeneous Group of Trajectories (*HGT*) which is filtered and re-sampled (step 3). Every position of each trajectory of the *HGT* are paired to a reference trajectory using Fréchet matching [5] to produce clouds of matched positions. A statistical analysis of these clouds gives the median trajectory of the *HGT* and spatio-temporal normality bounds (step 4) which are combined together to define the spatio-temporal pattern of the *HGT*. Every trajectory of the *HGT* is also compared to this pattern in order to compute similarity measurement statistics such as maximum or mean spatial and temporal distance. Finally, this pattern and its statistics are stored in a knowledge database (step 5). These spatio-temporal patterns can be used either for geovisual analysis or to qualify in real-time ship positions and trajectories. In the latter approach, every newly acquired position is matched to a spatio-temporal pattern. The selected spatio-temporal pattern is used to qualify the position and the trajectory (step 6) using fuzzy logic to obtain a combined spatio-temporal similarity rating and visualise it (step 7).

2 Introducing Geovisualisation for Anomaly Detection

Traditional data mining techniques can be aided by visual analysis tools to support human reasoning through interactive visual interfaces [1]. The term geovisual analytics represents this process in the spatial domain. In many cases, geovisual analysis can detect outliers or unusual behaviour which data-mining approaches miss. To support maritime anomaly detection, we are investigating the use of a geovisual analysis environment based on the space-time cube [4], in which, the x and y planes represent the spatial context while the z plane represents the temporal component. Trajectories are represented as a 1-dimensional mathematical piece-wise linear curve in the three-dimensional space, forming a 3D polyline in a space-time cube. Figure 2 shows an example of the visual environment displaying vessel movements for a particular route. Vessels which are not following the route in terms of progress (temporal positioning) and geographic locations can be visually identified. For example, it is clear to see a trajectory (in black) which is not following the typical route of other vessels in terms of its progress through the space.

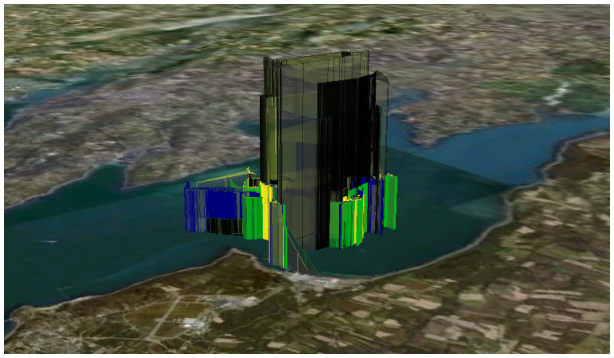


Figure 2: Space-time cube showing the path of vessels with one abnormal trajectory

When the quantity of trajectories becomes larger, visually identifying anomalies can be difficult. As a result, it is necessary to combine the space-time cube visualisation with the clustering and data mining approaches described above in order to visualise the most salient patterns. This assists human cognition while still providing appropriate tools to explore the spatial data.[1]

References

[1] G. Andrienko, N. Andrienko, J. Dykes, M. Kraak, and H. Schumann. Geova(t) - geospatial visual analytics: Focus on time. *International Journal of Geographical Information Science*, 24(10):1453–1457, 2010.

[2] F. Bertrand, A. Bouju, C. Claramunt, T. Devogele, and C. Ray. Web architecture for monitoring and visualizing mobile objects in maritime contexts. In *Proceedings of the 7th international conference on Web and wireless geographical information systems*, pages 94–105. Springer-Verlag, 2007.

[3] L. Etienne, T. Devogele, and A. Bouju. Spatio-temporal trajectory analysis of mobile objects following the same itinerary. In *Proceedings of the International Symposium on Spatial Data Handling (SDH)*, page 6, 2010.

[4] T. Hägerstrand. What about people in regional science? *Papers in Regional Science*, 24(1):6–21, 1970.

[5] A. Mascret, T. Devogele, I. Berre, and A. Hénaff. Coastline matching process based on the discrete fréchet distance. *Progress in Spatial Data Handling*, pages 383–400, 2006.

[1]Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan and by a grant (Ulysses programme) from the Irish Research Council for Science, Engineering and Technology (IRCSET) and the French ministry of foreign affairs. The authors gratefully acknowledge this support.

Maritime Anomaly Detection using Stochastic Outlier Selection

Jeroen H.M. Janssens, Eric O. Postma, and Jaap H.J. van den Herik
Tilburg center for Cognition and Communication, Tilburg University, Warandelaan 2, 5037 AB,
Tilburg, The Netherlands, {j.h.m.janssens,e.o.postma,h.j.vdnherik}@uvt.nl

We consider the challenge of detecting maritime anomalies in an automated fashion. Previous work demonstrated that the Stochastic Outlier Selection (SOS) algorithm has a superior performance on synthetic and UCI benchmark data sets [3]. However, when applying SOS to the complex maritime domain, the real challenge is to find an appropriate data representation. Our preliminary results show that maritime anomalies may be detected with SOS. In Section 1, we briefly describe the input to the SOS algorithm, namely a dissimilarity matrix of patches. Then, we explain the SOS algorithm using a thought-experiment (Section 2). Subsequently, we describe the application of SOS in a maritime use case (Section 3). Finally, in Section 4 we give our conclusions and provide directions for future research.

1 A dissimilarity matrix of patches

Our data consists of AIS vessel trajectories. To avoid comparing trajectories of different duration, we use from each vessel trajectory the last, say, thirty minutes. We refer to such a subsection of fixed duration as a *patch*. Our goal is to compute an outlier probability for each patch, so that we have an indication of whether the corresponding vessel is an anomaly or not. Given a measure of dissimilarity between pairs of patches (e.g., the alignment measure described in [1]), we may compute a so-called dissimilarity matrix, where the value in row a and column b represents the dissimilarity between patch a and b. When two patches are exactly alike, then their dissimilarity is zero. We require that the outlier outlier probability of a patch increases as it becomes less similar to the patches of the other vessels. For example, if one patch describes a u-turn whereas all the other patches follow a straight line, then this one patch is dissimilar to the other patches, therefor an outlier, and the corresponding vessel may also be an anomaly.[1]

2 The Stochastic Outlier Selection algorithm

The following description of the SOS algorithm is based on a thought experiment. A more technical description may be found in [3]. Please note that the maritime terms used in this description are chosen for explanation purposes only, and that the SOS algorithm is not bound to any domain.

Imagine, that to each vessel sailing on the North Sea, it appears (due to a hardware failure of some sort) as if they are simultaneously receiving distress calls from all other vessels. A vessel can only help one other vessel, so the vessel is forced to make a choice. We let the decision which vessel to help, base on the *willingness* to help the other vessel, which we define as a decreasing function of the dissimilarity value (cf. Section 1) between the two patches corresponding to the helping vessel and the other vessel. The vessel has some degree of willingness to all other vessels, just as it has some degree of similarity to all other vessels. This *willingness distribution*, causes stochasticity, or randomness, in its decision. All vessels simultaneously and independently decide which vessel to help, and those vessels which are not

[†]This work has been carried out as part of the Poseidon project under the responsibility of the Embedded Systems Institute, The Netherlands. This project is partially supported by the Dutch Ministry of Economic Affairs under the BSIK03021 program.

[1]We define an *anomaly* as an entity (e.g., vessel) in the real world that is considered to be abnormal (as judged by a domain expert), and we define an *outlier* as a data point (e.g., patch) in a data set that is not inside a cluster or high-density region (as judged by an algorithm). It is possible that an anomaly in the real world is not an outlier in the data set (and vice versa).

being helped by any other vessel, are considered to be outliers. To obtain the *probability* that a certain vessel is an outlier, we repeat the hypothetical decision process infinitely many times. The ratio of that vessel not being helped is the outlier probability.

3 SOS applied to the maritime domain

The Dutch part of the North Sea has 17 official sea lanes that guide large amounts of vessel traffic such as tankers and cargo ships. Operators at the Dutch coastguard are particularly attentive to these sea lanes (personal communication with maritime domain experts). Typical types of anomalies in these sea lanes include vessels that (1) suddenly stop, (2) sail the wrong way, and (3) perform u-turns.

The SOS algorithm as it is described in Section 2, computes an outlier probability for all patches at a certain time step. Since the data in the maritime domain are time series, repeated calculation of the outlier probabilities is required. To this end, we have extended SOS to a one-class classifier such that it has a training and testing phase. In this use case, SOS was trained with a data set that consisted of thirty-minute patches that were collected from a period of six days. The seventh day was used for testing.

The dissimilarity measure we have employed is based on the alignment measure as described in [1], with the addition that patches that are not in the same sea lane have an increased dissimilarity. This ensures that a "model of normality" is created per sea lane. We only include patches of tankers and cargo ships, as these are of most interest to the operator. Twenty-five artificial anomalies (belonging to the types described above) were created and inserted in the data by maritime domain experts using Presto [2]. These anomalies were used to evaluate our approach.

Our preliminary results show that most of the 25 artificial anomalies are selected as outliers by SOS, which is promising. We have observed two possible issues why some anomalies are not selected. Firstly, some sea lanes describe a very small area causing insufficient patches to be collected during training. Secondly, the original AIS data already contains some vessels that are anomalous, causing the artificial anomalies not be outliers. The first issue may be solved by having a separate dissimilarity matrix (and SOS algorithm) per sea lane, enabling easy stratification of the training data set. The second issue may be solved by cleaning up the training data by discarding patches whose outlier probability is too high.

4 Conclusion and future work

From our preliminary results, we may conclude that the SOS algorithm has potential for the maritime domain. We have learned that domain experts play an invaluable role in the process of finding an appropriate data representation, and evaluating the approach. In future research we intent to: (1) solve the two issues mentioned in the previous section, (2) apply our approach to other types of anomalies, and (3) compare the SOS algorithm to existing outlier selection algorithms in the maritime domain.

References

[1] G.K.D. de Vries, W.R. van Hage, and M.W. van Someren. Comparing vessel trajectories using geographical domain knowledge and alignments. In Fan. W, W. Hsu, G.I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, editors, *ICDM Workshops*, pages 209–216. IEEE Computer Society, 2010.

[2] J.H.M. Janssens, H. Hiemstra, and E.O. Postma. Creating artificial vessel trajectories with Presto. In *Proceedings of the 22nd Benelux Conference on Artificial Intelligence (BNAIC 2010)*, Luxembourg, Luxembourg, 2010.

[3] J.H.M. Janssens, E.O. Postma, and H.J. Van den Herik. Unsupervised Outlier Selection with Pairwise Affinities. In *SNN Symposium: Intelligent Machines*, Nijmegen, The Netherlands, 2010.

Maritime Route Anomaly Detection

Richard O. Lane
QinetiQ
Malvern, Worcestershire, UK
rlane1@QinetiQ.com

Ships involved in commercial activities tend to follow set patterns of behaviour depending on the business in which they are engaged. If a ship exhibits anomalous behaviour, this could indicate it is being used for illicit activities. With the wide availability of automatic identification system (AIS) data, as well as other non-cooperative sensors such as radar, it is now possible to detect some of these patterns of behaviour. Monitoring the possible threat posed by the worldwide movement of tens of thousands of ships between thousands of ports, however, requires efficient and robust automatic data processing to create a priority list for further investigation.

It is advantageous for ships to travel by the most economical route between two points on the ocean surface, which is often the shortest route defined by segments of great circles. Constraints on a ship's movement consist of land masses, depth of water, traffic separation schemes, weather, and exclusion zones. The constraints result in ship tracks following certain patterns. One exemplar existing technique models ship tracks using a Gaussian mixture model (GMM) while another uses a kernel density estimator (KDE). These are compared in [2]. Dividing the ocean into a number of grid cells, with a density sufficient for a specified accuracy, has the disadvantage that several million cells and connections between cells would be required to model global shipping movements.

This paper proposes an alternative model for the movement of ships at sea and uses a network of discrete nodes placed at route decision points relating to constraints, and branches to connect the nodes. The network model is sparse, being branch rich and node light. This has computational advantages since the speed of optimal route algorithms are dominated by the number of nodes rather than the number of branches. As the ultimate aim is to produce a global network, calculation time is an important issue for both network preparation and analysis. A node-sparse network might only need 20,000 nodes and have journeys that involve one to ten branches rather than hundreds. This model was first presented in [3] and [1] and is expanded here.

In the network design, each landmass in the world is represented by a closed polygon using coastline data. There are three types of node located on the coast: ports, convex hull coastal points, and other coastal points required to define journeys to ports. Offshore nodes are added for destinations such as oil rigs, for defining the edge of restricted routes such as traffic separation schemes, and for observed set routes not constrained by land. Great circle branches are generated between pairs of nodes that are not impeded by any intervening landmass. In addition, branches expected to be of no practical use in route planning are excluded. Once the nodes and branches have been established, Dijkstra's algorithm is used to pre-calculate optimal routes and costs between ports in the network. The baseline cost function for each branch is its length, but could be augmented by canal tolls or other branch specific features. Pre-calculation of optimal routes allows subsequent analysis of ships' routes to be sped up considerably.

A ship's journey begins and ends at a port and is assumed to follow a route that passes through or close to a sequence of intermediate network nodes. A complete journey J can be described by its constituent branches between those nodes such that $J = b_1, ..., b_n$ where the b_i represent branches. In practice, it is necessary to analyze incomplete journeys to assess whether the ship's behaviour has been anomalous or not. An incomplete journey is made up of two parts: the macro part and the micro part. The macro part consists of all branches traversed before the previous network node. The micro part consists of the route since the previous network node. It is useful to divide the route in this way as the macro route can be described with reference to the pre-defined network, whereas the micro route, having no local nodes with which to identify, has to use a more precise frame of reference.

The underlying requirement is to calculate, for every possible destination D_i, the probability that the true destination is D_i, conditional on having observed the route so far. This can be expressed as $P(D_i|R_M,R_m)$ where R_M represents the macro route and R_m represents the micro route. Using Bayes' theorem and making the prior assumption that all ports are equally likely destinations, it can be shown that $P(D_i|R_M,R_m) \propto P(R_M|D_i)P(R_m|D_i)$. $P(R_M|D_i)$ is calculated from the cost of the route taken compared to the optimal cost. $P(R_m|D_i)$ is calculated by comparing the ship's heading at each measurement point with the heading of candidate branches. Details of these calculations are given in [3].

After every new observation of a ship's position, conditional destination probabilities are updated for every destination in the network. These probabilities can be used to assess the following two hypotheses: H_0, the ship is travelling to its stated destination; and H_1, the ship is travelling somewhere else. Implicit in the null hypothesis H_0 is the assumption that if a ship is travelling to its stated destination, then it will do so by a route that does not incur unreasonable cost. The anomaly statistic of interest is $P(H_0|r_t)$ where r_t is the complete route covered by the ship up to time t. Using Bayes' rule this can be written as

$$A_t \equiv P(H_0|r_t) = \frac{P(r_t|H_0)P(H_0)}{P(r_t|H_0)P(H_0)+P(r_t|H_1)P(H_1)}.$$

Prior values for $P(H_0)$ and $P(H_1)$ are required. If it is observed that over a representative sample of ships' journeys, a proportion q of them result in a ship travelling to the stated destination D_s, then $P(H_0)$ could be set to q and $P(H_1)$ set to $1-q$. In this work, q has been set to 0.999. Since H_1 includes every destination that is not the stated one this covers $n-1$ destinations, where n is the total number of possible destinations. Hence the probability that the ship is travelling to a specific non-stated destination is $(1-q)/(n-1)$. For the case where there is no knowledge of the expected destination, the use of this anomaly statistic can be extended to a series of null hypotheses for each possible destination D_i to determine at each observation, which of the destinations appear feasible and which appear anomalous. At the start of a journey all destinations will appear feasible (i.e. the $A_{t,i}$ value for each hypothesized destination D_i will be close to unity) but they will gradually be whittled down as the journey progresses and the route becomes clearer. If at any point of the journey, every possible destination has appeared anomalous at some previous point (including the current point) then there remains no feasible place for the ship to travel to and it should be flagged up as anomalous. In other words, $A_{t,i}$ has fallen below a defined threshold for every destination D_i at some time t (t is not constrained to be the same for each destination).

A number of issues relating to real data have been required to be addressed. These include: associating an area rather than a point with the notion of a port; modelling ships that give headlands a wide berth; and efficiently parameterising routes using only those nodes that are the sufficient statistics for a route planner.

Using the above processes, the deviation from standard route algorithm has been applied to measured data of several hundred ships. Results of the algorithm show that it is indeed able to detect ship tracks that do not conform to expected routes between ports.

References

[1] R. O. Lane, D. A. Nevell, S. D. Hayward, and T. W. Beaney. Maritime anomaly detection and threat assessment. *13th Int. Conf. on Information Fusion, Edinburgh, UK*, July 2010.

[2] R. Laxhammar, G. Falkman, and E. Sviestins. Anomaly detection in sea traffic - a comparison of the gaussian mixture model and kernel density estimators. *12th Int. Conf. on Information Fusion, Seattle, WA*, July 2009.

[3] D. A. Nevell. Anomaly detection in white shipping. *Mathematics in Defence 2009, Farnborough, UK*, November 2009.

Applying V-Analytics to AIS Data

Gennady Andrienko
Fraunhofer Institute IAIS
53754 Sankt Augustin, Germany
http://geoanalytics.net/and

Natalia Andrienko
Fraunhofer Institute IAIS
53754 Sankt Augustin, Germany
http://geoanalytics.net/and

1 Introduction

In this short paper we use V-Analytics [2] research prototype software (a.k.a. CommonGIS[3]) for detecting patterns in AIS data. The example dataset used in this case study contains trajectories of 6103 ships over 8 days in the North Sea, having in total 444K recorded positions. The data have been provided by MARIN, the Maritime Research Institute in the Netherlands. [1].

V-Analytics incorporates various visualization techniques, interactive tools, and computational methods for analyzing spatial, temporal, and spatio-temporal data. Among them are time-aware maps, space-time-cubes, other types of graphs and diagrams; interactive dynamic filtering of data according to their spatial, temporal, and thematic (attributive) components; computational procedures oriented to movement data such as clustering of trajectories[1], generalization and summarization[4], detecting interactions between moving objects[5], extracting various types of events from movement data, etc.

2 Analyzing vessel movement

Generalization and summarization [4] extracts major traffic flows from movement data. We use filtering to remove trajectories with large sinuosity and tortuosity and apply the method to the remaining trajectories (figure 1). The method performs spatially-bounded clustering of characteristic points of the trajectories, builds Voronoi polygons around the centroids of the clusters, and represents trajectories as sequences of moves between the visited polygons. The results of the summarization are flows (aggregated moves) between the polygons, which are shown on a map as arrows with the widths proportional to the traffic volumes. The flows can be animated over time, interactively filtered, and further analyzed.

For detecting particular events in the trajectories, such as potentially dangerous proximity of ships, drifting etc., we apply tools for computing various movement attributes and for finding interactions among moving objects [5] together with interactive visual interfaces for filtering segments of trajectories according to their attributes. The extracted patterns are visualized on a map, in a space-time cube (figure 2), and other displays.

Our current implementation is a Java application that uses Oracle database for data preprocessing and performs most of the analytical operations in RAM. We are now developing methods that integrate interactive analysis with database computations in Oracle and specialized moving object databases.

References

[1] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 3 –10, oct. 2009.

---

[1] The Netherland Coastguard collects data of shipping movements by radar coverage and AIS base stations. MARIN receives the fused data for use in safety assessment studies with respect to shipping. MARIN has anonymized a week of data for this research/experiment. The authors are grateful to Y.Koldenhof (MARIN) for preparing the tasks and providing feedback

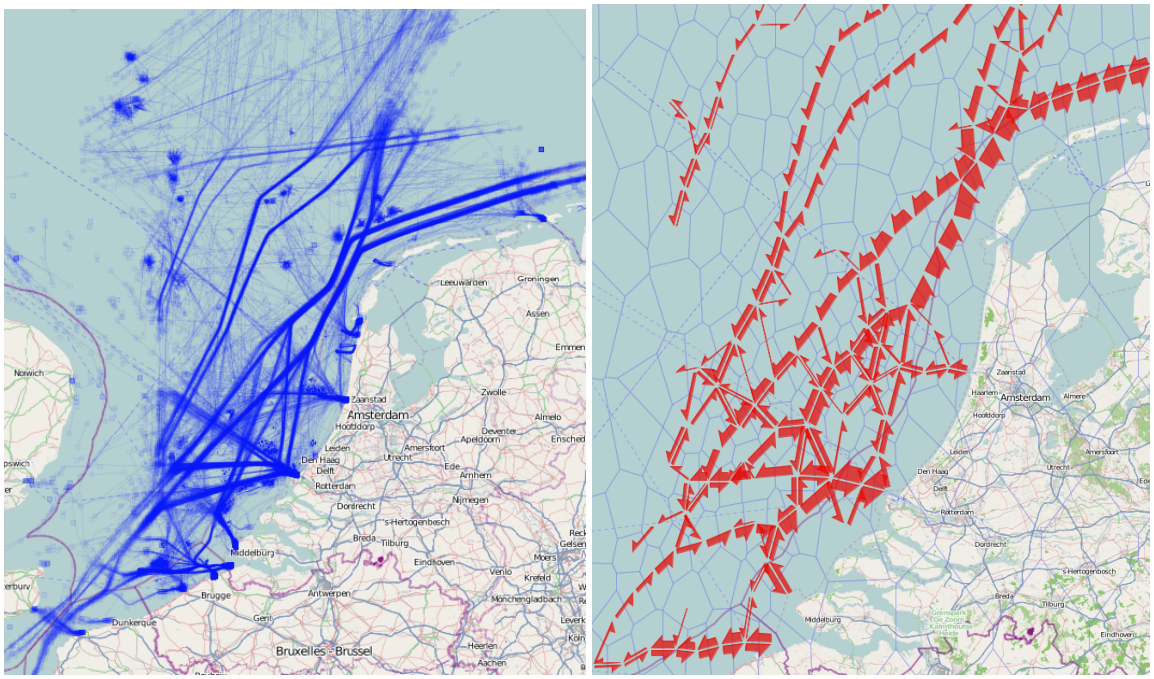


Figure 1: Left: trajectories are drawn with 95% transparency; right: flow map of the sea traffic

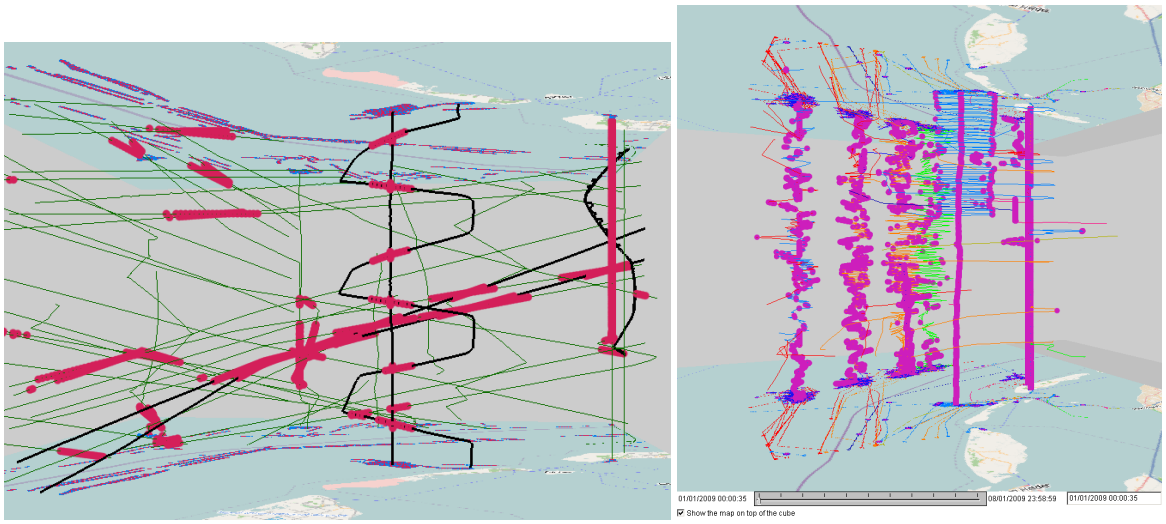


Figure 2: Space-time cubes show events of ship proximity (left) and drifting (right)

[2] Gennady Andrienko, Natalia Andrienko, and Stefan Wrobel. Visual analytics tools for analysis of movement data. *SIGKDD Explor. Newsl.*, 9:38–46, December 2007.

[3] Natalia Andrienko and Gennady Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer, 2006.

[4] Natalia Andrienko and Gennady Andrienko. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics*, 17:205–219, 2011.

[5] Daniel Orellana, Monica Wachowicz, Natalia Andrienko, and Gennady Andrienko. Uncovering interaction patterns in mobile outdoor gaming. *Advanced Geographic Information Systems and Web Services, International Conference on*, 0:177–182, 2009.

An Evaluation of Fractal/Velocity Pattern Extraction

René A. Enguehard
Dept. of Geography
Memorial University
of Newfoundland
St. John's, NL Canada
rene@computer.org

Rodolphe Devillers
Dept. of Geography
Memorial University
of Newfoundland
St. John's, NL Canada
rdeville@mun.ca

Orland Hoeber
Dept. of Computer Science
Memorial University
of Newfoundland
St. John's, NL Canada
hoeber@mun.ca

Abstract

Visual analysis of maritime tracking data has been shown to be a useful complement to automated approaches. However, detecting interesting or anomalous movement patterns can be a challenging task due to the size of spatio-temporal datasets. Datasets can include a large amount of uninteresting data, which hinders the task of pattern extraction. This paper presents a method by which uninteresting patterns can be filtered out, based on their velocity and fractal properties, simplifying the pattern extraction problem considerably. This technique is then compared to an automated statistical pattern extraction technique, called Behavioural Change Point Analysis, by a number of experts in the course of a field trial.

1 Introduction

Visual analysis of maritime tracking data has been shown to be a useful complement to automated approaches [6]. Detecting specific patterns within a data set can be challenging due to the large volume of data to either process or visualize [5]. Vehicle tracking data sets, such as those generated by Vessel Monitoring Systems (VMS) or Automated Identification Systems (AIS), often consist of hundreds of thousands of data points, with most of them not being part of any interesting patterns or behaviours. In these cases, the vast majority of the data is not interesting, so it can be very difficult for users to detect relevant patterns. By taking into account the velocity ranges of particular activities, as well as the fractal nature of the movements, these patterns can be detected and filtered out, if required.

Many approaches have been proposed for pattern extraction, from a multitude of fields, such as signal processing [2], finance [1] and biology [4]. For this study, we chose to compare an approach recently proposed[3] to one from the field of biology, Behavioural Change Point Analysis (BCPA) [4]. This technique is designed for use in animal tracking, statistically analyzing sets of velocities (or velocity components) in order to locate the points in the data where the observed behaviour changes. This is done by treating each candidate change point as the point on which to split the data set in half. These two halves are then tested for statistical difference, with the highest-likelihood point becoming the Most Likely Change Point (MLCP). The entire process is repeated using a moving window, yielding a set of potential change points within the full data set.

The data used in this project come from VMS units installed on-board fishing vessels operating off the East coast of Canada. The use of VMS is a condition for license by Fisheries and Oceans Canada (DFO), the Canadian government's fisheries regulation department. This is done to aid DFO in monitoring when and where captains fish, providing an indication of the distribution of fishing effort. The VMS units record the latitude and longitude of each vessel at a predetermined time-interval, usually one hour, and then transmit these data via a satellite link.

2 Fractal/Velocity Signatures

By taking into account both the velocity and the complexity of movements, it is possible to categorize movements into particular pattern classes. Doing so in an exploratory and interactive fashion is what we term fractal/velocity signature building. It is essentially an iterative process by which users develop a set of filter settings, leaving only the patterns they are interested in. These signatures can then be saved and applied to multiple similar data sets to allow for rapid extraction of the target patterns, and hence allow more efficient data analysis processes.

Signatures are made up of velocity and fractal dimension ranges. Velocity is estimated using linear distance between data points, divided by the time between data points. The fractal dimension, a parameter which estimates the complexity of a path, is estimated using a moving window. By varying the width of the moving window, the temporal range of the behaviour under investigation can be adjusted.

3 Evaluation

In order to validate the fractal/velocity signature approach, we compare our proposed technique with BCPA. While the main goal of both techniques is to aid in the detection of specific patterns, BCPA is an automated approach, whereas our work is interactive. Both approaches were shown to fisheries enforcement officers working for DFO. The primary task in their work is to detect irregular or illegal fishing activities, often using VMS datasets to identify such patterns. As such, they were very familiar with the pattern extraction activities both system support, and could therefore gauge the usefulness of one technique versus the other.

The evaluation proceeded by presenting the experts with the interactive interface first, on an individual basis, wherein they were able to develop signatures to match the various patterns they would be interested in within a pre-determined data set. They were then shown the output from the BCPA method, which they could visually explore. Finally, they were asked to rate each method in terms of usability and usefulness, as well as giving specific comments as to strength and weaknesses of each method. An overview of the approach, along with the results of this field trial will be discussed in this presentation.

References

[1] M. Dong and Z. Xu-Shen. Analyzing dividend events with neural network rule extraction. In *Proceedings of the International Joint Conference on Neural Networks*, pages 2854–2859, Portland, OR, USA, 2003. IEEE.

[2] B. Eisenstein and J. Fehlauer. Signal processing for feature extraction and pattern recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 749–752, Philadelphia, PA, USA, 1976. IEEE.

[3] R. Enguehard, R. Devillers, and O. Hoeber. Geovisualization of fishing vessel movement patterns using hybrid fractal/velocity signatures. In *Proceedings of GeoViz Hamburg*, 2011.

[4] E. Gurarie, R. D. Andrews, and K. L. Laidre. A novel method for identifying behavioural changes in animal movement data. *Ecology Letters*, 12:395–408, 2009.

[5] P. Lundblad, M. Jern, and C. Forsell. Voyage analysis applied to geovisual analytics. In *Proceedings of the 12th International Conference on Information Visualization*, pages 381–388, London, UK, 2008. IEEE.

[6] M. Riveiro and G. Falkman. Evaluating the usability of visualizations of normal behavioral models for analytical reasoning. In *Proceedings of the 7th International Conference on Computer Graphics, Imaging and Visualization*, pages 179–185, Sydney, Australia, 2010. IEEE Computer Society.

Incremental Stream Clustering for Anomaly Detection in Maritime Surveillance

Anders Holst
Swedish Institute of Computer Science
Sweden
aho@sics.se

Jan Ekman
Swedish Institute of Computer Science
Sweden
jan@sics.se

Abstract

We have designed a framework for Bayesian Statistical Anomaly Detection, called ISC, or Incremental Stream Clustering. It learns the normal situation incrementally, and can on the fly detect anomalous cases. When this happens, a new cluster can be created, so similar cases can be detected in the future. In this way, the framework performs incremental clustering, classification and anomaly detection at the same time. The framework will be used in a recently started project to support Maritime Domain Awareness.

1 Background

Anomaly detection is a growing area with more and more practical applications every day. It has been used for fraud detection and intrusion detection for a long time, but in later years the usage has exploded to all kind of domains, like surveillance, industrial system monitoring, epidemiology, etc. Recently it has also attracted much attention within maritime domain awareness. For an overview of different anomaly detection methods and applications, see e.g [2].

The approach taken in *Statistical anomaly detection* is to use data from (predominantly normal) previous situations to build a statistical model of what is normal. New situations are compared against that model, and are considered anomalous if they are too improbable to occur in that model.

To be useful for real world application, a statistical anomaly detector must fulfil several conditions:

- It must be able to handle a large number of input features, and several different normal situations.
- It must allow for training data to include realistic amounts of anomalous cases.
- It must be fast when dealing with large amounts of data, and at the same time robust in the light of very small amounts of training data for some situations or features.
- It must have a sufficiently small false alarm rate, while still being maximally sensitive to real anomalies.

In spite of the large amounts of different anomaly detection algorithms developed, most of them fails on one or more of the above requirements. We have developed a novel anomaly detection method based on Bayesian statistics, with the above properties. The strength of Bayesian statistics in this context is its ability to handle uncertainties of different types in a consistent way, including the uncertainties introduced by noisy or scarce training data, and as a consequence the possibility to control the false alarm rate while maintaining sensitivity.

The *Incremental Stream Clustering* framework (ISC) provides a combination of anomaly detection, clustering and classification [3, 4]. In the center there is our statistical method for anomaly detection based on Bayesian statistics. It can either be used batch-wise, i.e it is first trained on one data set, and then used to find anomalies in another, or on streaming data, i.e it both learns from and detects anomalies in new data as it arrives. In the latter case it can also be used for clustering: when a new sample arrives

which is considered anomalous, i.e. not coming from the same distribution as previously seen, then a new cluster is formed around that sample, in the form of a new anomaly detector which considers that sample as "normal". With the help of a human operator who inspects new anomalous samples, the new clusters may be given appropriate labels. In that way classification and anomaly detection is combined, and the system learns to classify more and more situations as it is used.

2 ISC for Maritime Domain Awareness

In the recently started project SADV, *Statistical Anomaly Detection and Visualization for Maritime Domain Awareness* [1], SICS will together with Saab, the Swedish Coast Guard, the Swedish Customs Service, the Swedish Armed Forces, and the Swedish Space Corporation, provide anomaly detection capabilities to the Swedish maritime surveillance platform *Sjöbasis*.

With an increasingly complex maritime situational picture, operators cannot rely on simply their eyes in order to detect anomalies or suspicious vessel behavior. Due to the sheer volume of incoming sensor data, it becomes both impractical to visualize and hard to manually detect the important aspect of the vessels on a nautical chart. Furthermore, several suspicious behaviors need not be practically observable by human operators, or may need to be observed over long periods of time to be detected.

The goal of the SADV project is to provide a significantly increased maritime domain awareness, by developing tools and methods for automatic extraction and highlighting of vessels with anomalous or suspicious behavior. The ISC framework will be at the core of this.

There are many platforms in use for maritime awareness and security. However, most of these systems has no support for automated anomaly detection. Those that do, have very limited anomaly detection capabilities, typically only considering simple kinematic features like speed and position.

Our approach is to combine input from different sources and actors, such as radar and other sensors, responder systems, customs documentation, and historical data over e.g. customers and vessels. We will then use the ISC framework for anomaly detection, to sieve through and extract entities deviating from expected behavior. Finally, this information will be presented to the operators and analysts in an intuitive and easily accessible way. The approach will make it possible to both detect short term anomalies based on kinematic and dynamic features, and long term anomalies based on more static and historical data like ports visited and type of transported cargo.

This approach will yield methods that assists the operators and analysts in real time to sieve through large amounts of incoming data, looking for complex patterns and integrating over long time spans, in order to find the vessels requiring attention. The methods are also, as described above, able to adapt automatically as the world changes; take feedback and learn from the operators; and are both sensitive to real anomalies and robust against false alarms.

References

[1] B. Bjurling, A. Holst, O. Ståhl, and A. Wallberg. Statistical anomaly detection and visualization. In *Proc. of TAMSEC 2010, First National Symposium on. Technology and Methodology for Security and Crisis Management*, 2010. Linköping, 27–28 October, 2010.

[2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009.

[3] J. Ekman and A. Holst. Incremental stream clustering and anomaly detection. SICS Technical Report T2008:01, Swedish Institute of Computer Science, Kista, Sweden, 2008.

[4] A. Holst and J. Ekman. Incremental stream clustering for anomaly detection and classification. In A. Kofod-Petersen, Fredrik Heintz, and Helge Langseth, editors, *Eleventh Scandinavian Conference on Artificial Intelligence, SCAI 2011*, pages 100–107. IOS Press, 2011. Trondheim, 24–26 May, 2011.

A Bayesian Network Approach to Maritime Situation Assessment

Yvonne Fischer
Institute for Anthropomatics
Karlsruhe Institute of Technology
yvonne.fischer@kit.edu

Joris IJsselmuiden
Fraunhofer Institute of Optronics,
System Technologies and Image Exploitation
joris.ijsselmuiden@iosb.fraunhofer.de

Abstract

This article presents a method for maritime situation assessment based on Bayesian networks. Our method is illustrated through an example scenario about illegal immigration.

1 Introduction

Our goal is to develop a maritime surveillance system that increases an operator's situation awareness. This can be done by guiding his focus of attention to the most suspicious activities that are currently going on. The representation of observed vessels in a wide maritime area can be done using a world model, which is a representation of entities in the real world [1]. The representatives in the world model are derived by observing the real world through sensors and analyzing the resulting sensor data with signal processing methods. The most common sensors are video, radar, and AIS. The resulting world model then serves as the basis for inference processes that generate hypotheses about what situations are taking place. These hypotheses can increase the operator's situation awareness and guide his or her focus of attention. Inference processes can also be used to generate action plans for the real world, for example in sensor deployment planning. Figure 1 shows our general system architecture and its information flows.

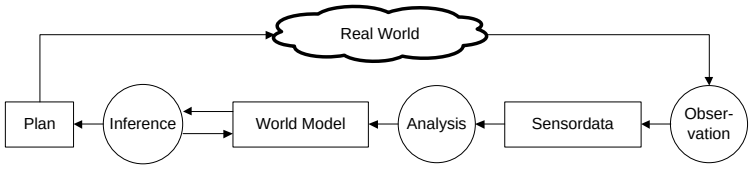


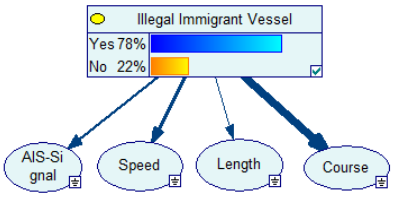
Figure 1: System architecture for maritime situation assessment

Example scenario. To demonstrate how hypotheses are inferred from sensor data, we now infer whether or not an observed vessel on the Mediterranean Sea is likely to hide illegal immigrants on board. A small wooden boat departs from the Tunisian coast, approaches the island of Lampedusa and finally lands at the beach. The vessel has the following characteristics: it sends no AIS-signal, it travels with around five knots, it is made of wood, it is about 15 meters long, and it takes a more or less straight course towards the island of Lampedusa.

2 Method

The Bayesian network [4] for the example scenario is depicted in Figure 2 (left). The root node shows the probability that a detected vessel is hiding illegal immigrants on board. The other nodes describe the vessel characteristics from the example scenario above. An expert determines

the conditional probabilities that are shown in Figure 2 (right). If all evidences for an illegal immigrant boat are set, the infered probability is 78%. It will never reach 100%, because based on the four characteristics, there is still a possibility that the boat is not hiding illegal imigrants. The thickness of the edges in Figure 2 (left) represent connection strengths. In this case, the course towards Lampedusa has the largest influence on the probability that a vessel is hiding illegal immigrants. Such a simple Bayesian networks and its probabilities can easily be defined by maritime surveillance experts. The result of this probabilistic situation assessment is a degree of belief of a modeled situation rather than a hard classification.



	Illegal Immigrant Vessel	
	Yes	No
no AIS	0,9	0,7
sends AIS	0,1	0,3
speed less than 10kn	0,8	0,5
speed more than 10kn	0,2	0,5
length shorter than 20m	0,6	0,4
length longer than 20m	0,4	0,6
Course towards Lampedusa	0,9	0,2
other Course	0,1	0,8

Figure 2: Bayesian network for illegal immigrant vessel and its assigned probabilities

3 Discussion

This article presents our ongoing work on a maritime surveillance system that increases the operator's situation awareness. Our approach is fundamentally different from learning statistical models. We believe that an expert system, without learning, is the best approach for complex, heterogeneous high-level domains. Our broader goal is to develop a general purpose framework for high-level situation understanding that can work with arbitrary sensor setups and operator goals. Our institutions are experienced in high-level situation understanding for maritime surveillance, indoor surveillance, smart rooms, household scenarios for humanoid robots, surveillance aboard military vehicles, and other applications. One example is our recent work in the SmartControlRoom lab [3]. Here, the described system architecture and knowledge formalization is used, combined with a first order logic engine containing expert knowledge. We are currently replacing this logic engine with a more advanced algorithm, using Fuzzy Metric Temporal Horn Logic [2]. This allows us to gracefully handle temporal conditions as well as both forms of fuzziness: uncertainty and vagueness. In our future work, we will perform thorough experimental evaluations of our Bayesian and fuzzy inference algorithms. We also aim to establish their common base as general purpose framework for high-level situation understanding.

References

[1] Y. Fischer and A. Bauer. Object-oriented sensor data fusion for wide maritime surveillance. In *Proceedings of the 2nd International Conference on Waterside Security*, 2010.

[2] J. Gonzalez, D. Rowe, J. Varona, and F. X. Roca. Understanding dynamic scenes based on human sequence evaluation. *Image and Vision Computing*, 27(10), 2009.

[3] J. IJsselmuiden and R. Stiefelhagen. Towards high-level human activity recognition through computer vision and temporal logic. In *Proceedings of the 33rd Annual German Conference on Advances in Artificial Intelligence*, 2010.

[4] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, 1988.

Maritime Anomaly Detection by Fusing Sensor Information and Intelligence

A.C. van den Broek, R.M. Neef
P. Hanckman, A.J.E. Smith
TNO
The Hague, The Netherlands
bert.vandenbroek@tno.nl

F. Bolderheij
NLDA/CAMS
Den Helder, The Netherlands
f.bolderheij@nlda.nl

Abstract

To improve situation awareness and threat detection capabilities in maritime scenarios the combination of sensor-based information with context information and intelligence from various sources is required. In the study the fusion and analysis for revealing anomalies and suspect from normal behavior are based on domain ontologies. A test bed allows the study of various exploitation and assessments techniques applied to these domain ontologies. Using an appropriate scenario we have simulated suspect and normal behaviour to test the applicability of these techniques.

1 Introduction

Because of global economic and socio-political changes, an increase of conflicts near the world's coastlines is anticipated. The littoral zone is characterized by intense regular vessel traffic. The conduct of Maritime Security Operations and Peace support Operations therefore means that navies have to control instead of dominate the sea requiring information superiority and adequate situation awareness. For the purpose to achieve information superiority a research program at the Netherlands Organisation for Applied Scientific Research (TNO), in collaboration with the Royal Netherlands Navy (RNLN), has started aiming at improving maritime situation awareness.

Security operations are often characterized by controlling large areas with a limited number of assets. One of the main operational tasks is therefore to direct assets timely to the right position. For this purpose the so-called *Recognized Maritime Picture* (RMP) is used containing tracks of vessels which have been evaluated with respect to the activity and intentions of the vessels. The evaluation requires recognition of objects present in the scene, their interaction with the environment, and fusion with a priori information. Analyses of the recognized processes in the RMP allow forecasting of future activities and situations (i.e. situation awareness [1]) including those which may be threatening.

2 Methods for achieving Situation Awareness

In our case the vessel is the object of interest and the requested information for the RMP is the *vessel intent*: e.g. trade or fishery (normal process) or smuggling, piracy (anomaly/ threatening process). Situation and threat awareness should be achieved by combining information describing the current situation expressed by *observables*, and a priori knowledge expressed by *intent a priori information* with signatures of threatening and normal processes expressed by *indicators*.

The fusion process requires that *observables*, *indicators* and *intent a priori information* and their interrelations are represented in a meaningful manner and readily accessible to the processing system. For this purpose we need a context model, i.e. an information model that captures the context consisting of concepts and relationships that are relevant in our application domain and that ensures consistency and a common vocabulary.

We create such a context model via ontologies (see 1). We distinguish two types of ontologies: content ontologies, and situation ontologies [3, 2]. Content ontologies capture elements of interest in the application domain, such as known types of vessels and ports, and other domain-specific concepts. Situation ontologies capture a situation or series of states in the application space using concepts from the content ontologies. The situation ontologies form the constituents for our search patterns of interest, e.g. the specific intents that we want the system to recognize.

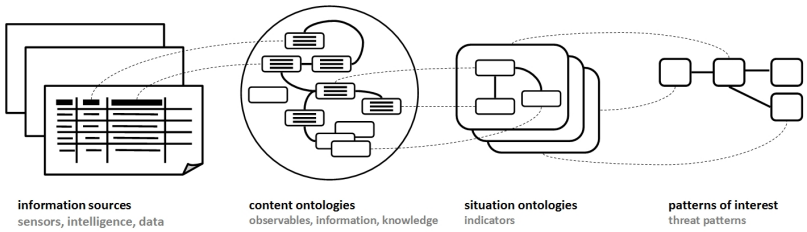


Figure 1: The elements of the context model. The situation ontologies form the basis for the definition of patterns of interest that may be used for threat assessment.

Situation recognition requires that relevant relationships between *indicators*, *observables* and *intent a priori information* are observed and assessed. The content ontologies provide semantic relationships; the situation ontologies define patterns that express how these semantic relationships lead to the *vessel intent*.

The goal is to provide alerts to suspected vessel behaviours, which can be expressed in patterns of interest, grounded in the context model. Threat alerts can be produced by mapping threat patterns on the available data. By assessing the semantic relationships between data, we can conclude whether a threat alert should be issued. Examples for assessment are probabilistic models such as Bayesian belief networks. The Bayesian approach is appropriate when the information obtained is uncertain which is often true in complex situations where hostile intent has to be inferred in the midst of overwhelming normal activities. Another approach in less complex situations is the use of rules and decision trees for obtaining actionable information.

Above-mentioned procedures and techniques are to be implemented in a situation awareness support system. To test the system we use a maritime scenario in a sea strait with dense trafficking. We use J-ROADS [4] for the simulation of the scenario and the modelling of sensors. Using the system, scenario and simulation we can determine the applicability of the various techniques: ontologies, Bayesian belief networks, decision trees and rules. At the time of writing this paper we are producing results which have to be evaluated before conclusions can be drawn.

References

[1] M.R. Endsley. Toward a theory of situation awareness in dynamic systems. In *Human Factors*, 37, pages 32–64, 1995.

[2] M.M. Kokar, C. J. Matheus, and K. Baclawski. Ontology-based situation awareness. In *Information Fusion*, volume 10, pages 83–98, 2009.

[3] J. Llinas, C. Bowman, G. Rogova, A. Steinberg, E. Waltz, and F. White. Revisiting the jdl data fusion model ii. In *7th International Conference on Information Fusion*, Stockholm, Sweden, 2004.

[4] W. van der Wiel. J-roads air defence simulation support during the 2006 jpow ix missile defence exercise. *NATO RTO, MSG-045-20*, 2006.

Web-based Geographical Visualization of Container Itineraries

Tatyana Dimitrova, Evangelos Kotsakis
Joint Research Centre, Institute for the Protection and Security of Citizen, Ispra, Italy
tatyana.dimitrova@jrc.ec.europa.eu evangelos.kotsakis@jrc.ec.europa.eu

Abstract

Around 90% of the world cargo is transported in maritime containers, but only around 2% are physically inspected. This opens the possibility for illicit activities. A viable solution is to control containerized cargo through information-based risk analysis. Container route-based analysis has been considered a key factor in identifying potentially suspicious consignments. Essential part of itinerary analysis is the geographical visualization of the itinerary. In the present paper, we present initial work of a web-based system's realization for interactive geographical visualization of container itinerary.

1 Introduction

Visualization and interaction techniques are widely recognized to be very powerful in anomaly detection [1]. In addition, visualization methods take advantage of human abilities to perceive patterns and to interpret them, which can be critical in complex situations, especially when the dataset is multi-dimensional, massive and dynamic. Such kind of data is the database of Contraffic (CDB), a system used to gather and analyse maritime container movements and screens the data on global maritime container movements to detect potentially suspicious consignments. The experience shows, that when the dataset is really huge, it is really difficult to assess what kind of data drilling techniques could be used for transforming the data. In such cases, the visualization of the data could help significantly by giving an initial general view of container behaviour. In certain cases, this can also indicate abnormal movements, i.e. circles in the container itinerary or unnecessary transshipments.

2 Review of existing solutions

There are two kinds of web solutions for container tracking. The first one is by the official cargo's web sites. Some of them offer the possibility for container/vessel tracking and geographical visualization of the current position on a map. The second one is by third parties web sites offering container and vessel tracking, i.e. http://www.marinetraffic.com, http://www.searates.com. The common disadvantage of these web sites is the very short history of tracking, which doesn't give any possibility for deep analysis of the container itinerary.

3 Contraffic Database

CDB contains currently more than 660 million of container events, which is around 30% of the current worldwide activity, and it is referencing to more than 12 million containers, which means around 300 000 new records are collected every day [2]. Each event describes an operation done on a container (e.g. load, discharge, transshipment) and it is related to additional information, including location, loading status of the container, carrier, type of the event, container identity number, time and date, vessel, bill of lading. The data for container movements is gathered automatically by heterogeneous semi-structured data sources [3] (on-line open public sources), where most of the information is not presented according to a format structure, but it is given as a textual description. In addition, different cargos use different formats for presenting the data. To make a homogeneous data representation, the gathered data goes through

a cleaning process. This process focuses on the codification of events and locations. The codification of the locations uses the UN/LOCODE data source, which is the United Nations Code for Trade and Transport Locations. The process of cleaning is done automatically when full pattern matching occurs, and manually by special designed software when no full matching occurs.

4 Realization of the web-based application

A Web-based application has been developed in PHP that allows the user to visualize the itineraries of a container over time using the container movements stored in CDB. The application visualizes the itinerary of a container using the container ISO identification number and a time window. The date fields have been realized by jQuery (JavaScript library) like data range pickers. Oracle connection to CDB is established and a dynamically created SQL query is executed. The query results contain detailed information about the route of this container over the requested time. The itinerary is visualized by Google Maps API and Google Earth API. Both visualizations are connected with each other, so move or zoom operations in one of them creates a corresponding effect to the other. The initial location is visualized by red marker, the final by blue and middle locations by green. Each of the connections between the locations is shown by lines of different color. For each of the locations an info-window was created giving detailed information about the event, dates of arrival and departure, vessel name and location details.

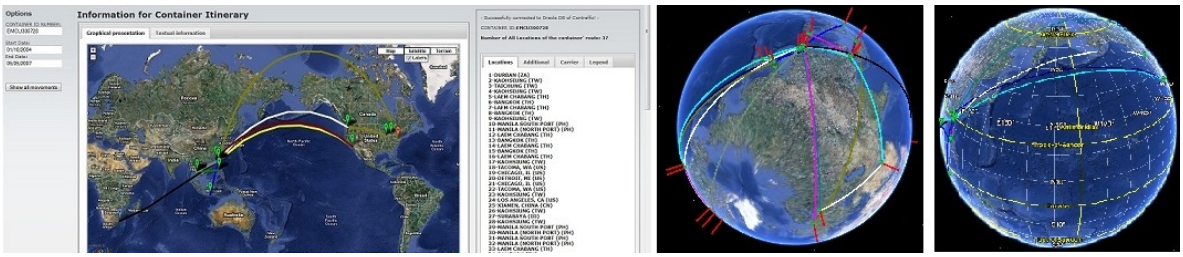


Figure 1: Geographical visualization of container itinerary

5 Conclusion and future work

The graphical visualization of container itinerary is an attractive way for analyzing container behaviour considering its historical trajectory. It could be useful as well for observing and investigating certain container movement and understanding the regular trajectories. The project needs to be enlarged by adding additional details for carrier routes and time schedule, Automatic Identification System data, improvement of the visualization of the real trajectory between pairs of ports, time graphs of the container voyage, vessel/bill of lading tracking, tracking by date and by port.

References

[1] P. Compieta, S. Di Martino, M. Bertolotto, F. Ferrucci, and T. Kechadi. Exploratory spatio-temporal data mining and visualization. Journal of Visual Languages Computing, 18(3):255–279, 2007.

[2] A. Donati, E. Kotsakis, A. Tsois, F. Rios, M. Zanzi, A.Varfis, T. Barbas, and J. Perdigao. Overview of the Contraffic System. Technical report, JRC, 2007.

[3] P. Villa, E. Camossi, A Description Logic Approach to Discover Suspicious Itineraries from Maritime Container Trajectories, GeoSpatial Semantics, Lecture Notes in Computer Science, Vol.6631, 2011.

Finding Fraud in Health Insurance Data with Two-Layer Outlier Detection Approach

R.M. Konijn, W. Kowalczyk

Abstract

Conventional techniques for detecting outliers address the problem of finding isolated observations that significantly differ from other observations that are stored in a database. For example, in the context of health insurance, one might be interested in finding unusual claims concerning prescribed medicines. Here, each claim may contain information on the prescribed drug (its code), volume (e.g., the number of pills and their weight), dosing and the price. Finding outliers in such data can be used for identifying fraud. However, when searching for fraud, it is more important to analyse data not on the level of single records, but on the level of single patients, pharmacies or GP's.

In this paper we present a novel approach for finding outliers in such hierarchical data. The novelty of our approach is to combine standard techniques for measuring outlierness of single records with conventional methods to aggregate these measurements, in order to detect outliers in entities that are higher in the hierarchy. We applied this method to a set of about 40 million records from a health insurance company to identify suspicious pharmacies.

1 Introduction

The inspiration for this paper comes from a real life fraud detection problem in health insurance. The goal of fraud detection in this context is to identify the most suspicious pharmacies that could possibly be involved in fraudulent activities, rather than identifying single claims that are suspicious. The main reason for not focusing on single outliers, is that recovering money from single claims is costly, and that it can harm the relationship between an insurance company and the involved pharmacy, especially in the case of false positives. On the other hand, if the insurance company can detect substantial fraud linked to multiple claims of the same pharmacy, this business relationship is no longer so important and a vigorous money recovery action can follow.

In contrast to typical approaches for finding single outliers, [2], we propose a novel method for finding *groups* of outlying records that belong to the same class. Our method was successfully applied to a large set of health insurance claims, helping to identify several pharmacies involved in fraudulent behaviour.

2 Our Approach

Our method for detecting group outliers works in two stages. In the first stage we calculate outlier scores for single records. We use here classical methods for outlier detection that are based on distance measures, or density estimation, [1].

Next, we calculate a statistic to measure the 'outlierness' of each groups of records, where groups form logical entities (in our case pharmacies). We apply four different statistics that are used to define the final outlier score of these entities: (1) a rank based statistic, (2) a weighted rank based statistic, (3) a statistic based on the binomial distribution, and (4) a statistic that is based on the mean of the outlier score. These statistics can be applied in different situations for different outlier scores.

The statistics can be computed over different segments of the data to obtain the final score. Extra information about outlying entities is obtained by constructing so-called *fraud sets*: sets of suspicious claims from the same pharmacy. A fraud set is a set of outlying records that should be removed from the whole set in order to make it "normal" again. Another, very useful instrument for displaying fraud

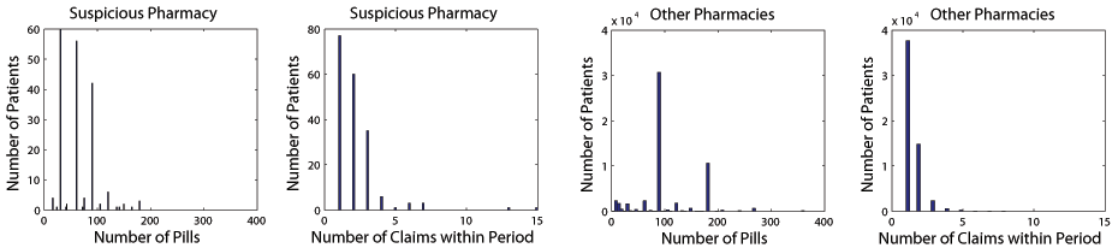


Figure 1: Histograms of the "number of pills prescribed" and the "number of claims" variables for the drug type *Aspirin*, both measured during the same period, over all pharmacies. The two histograms on the left show the distribution of the same variables calculated for the suspected pharmacy. From these graphs it can be concluded that these distributions are different. The number of pills is much lower on average, while the number of claims is higher. This is a clear indication of *unbundling* fraud.

evidence is an *fraud plot*: a plot of fraud amount versus outlier score of all records in the fraud set. Here, the *fraud amount* is defined as the total amount of money that is involved in the observations that are in the fraud set. The fraud plot can be used by fraud investigators to decide whether they should investigate the most likely fraud cases, or to focus on cases that are less suspicious, but involve high amounts of money.

3 Results

We applied our method to detect fraud in a set of 40 million claims from various pharmacies. For three different types of problems, we first calculated a single point score and then we identified "outlying pharmacies" by using one of the four statistics mentioned earlier.

A common type of fraud in health insurance is *unbundling*: a practice of breaking what should be a single charge into many smaller charges. Two other common types of fraud are: delivering more units than stated on the prescription (and thus charging more money), and charging money for drugs that have never been delivered. We identified these types of fraud simultaneously by mining "Local Outliers" at the patient level and then calculating an aggregated score per pharmacy by means of the weighted rank statistic. The most suspicious pharmacies do indeed have strange claim behaviour.

We zoom in at the top outlier and plot the distributions of two variables, see Figure 1. It is clear that the distributions of these variables are very different from the distributions of other pharmacies.

References

[1] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, 2000.

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41:15:1–15:58, July 2009.

Semantic-based Anomalous Pattern Detection from Maritime Trajectories

P. Villa and E. Camossi
European Commission Joint Research Centre
Institute for the Protection and Security of Citizen
Ispra, Varese, Italy
paola.villa@jrc.ec.europa.eu elena.camossi@jrc.ec.europa.eu

1 Introduction

In the field of maritime security, the analysis of moving object trajectories is a novel approach for fighting commercial frauds [1]. The resulting *Route-based Risk Indicators (RRIs)* may be used to evaluate the trajectories of cargos, ships, and containers, targeting high-risk consignments of goods and proceeding with physical inspections only when necessary. RRIs analyse the route followed by a container, relying on spatial information such as the ports where a it has been loaded, discharged, or transshipped. Their evaluation may complement that of traditional risk factors such as the name of the consignee, the carrier, the value of transported goods, etc. Despite the adoption of RRIs requires a remarkable computational effort, mainly because of the inherent complexity of spatial information, risk analysis may benefits of more effective results.

In our work, we propose *Semantic Route-based Risk Indicators (SRIs)* to enhance RRIs with the exploitation of the semantics of the domain of containerised cargos. Specifically, we develop the Maritime Container Ontology (MCO) [2] to model the explicit domain features of maritime containers, and relying on it we formalise SRIs as anomalous itinerary patterns, defined in the ontology as logical axioms. Thanks to this logic-based representation, we can reason on container itineraries inferring which of them have a suspicious behaviour.

Our work relies on previous investigations in the area of Geographical Information Science for the exploitation of the semantics hidden in trajectories. In [3], the logic-based representation of trajectories has been used to reason on touristic itineraries. Moreover, in [4] the semantics embedded in the geographical is exploited information to obtain a higher abstraction level.

2 Maritime Container Ontology (MCO)

The MCO is a knowledge base defined in OWL 2 Web Ontology Language (see `http://www.w3.org/`) that describes the domain of containerised cargos, and it is the foundation for SRIs development. The domain of maritime containers is very complex, but at the same time it offers a rich semantics to characterise container trajectories. In MCO we focus in particular on the *Events* that can occur to a container, which identify every handling activity performed during the container shipping (e.g., discharged, transshipped). Time-ordered sequences of container events are segmented into *Container Itineraries*, which embeds the trajectories that containers follow during a consignment. Similarly, the one-step itineraries that link pairs of ports, in the MCO are instances of the *Vessel Routes* concept.

In our experimental evaluation we considered a dataset of 16 millions container events, collected between 2009 and 2010, which have been processed to build the corresponding container itineraries. Such dataset has been uploaded into the MCO through an API developed with Jena (see `http://jena.sourceforge.net/`).

3 Semantic route-based risk indicators (SRIs)

SRIs crosscheck container itineraries with vessel routes to detect anomalous behaviours, and are implemented as logical axioms in the MCO, namely *SRI-axioms*. Compared to work adopting similar approaches for the evaluation of moving object trajectories [3], SRI-axioms are quite complex, because each axiom usually evaluates a container itinerary and several vessel routes, partially overlapping with it. This is because each container shipping is usually accomplished by more vessels, to which the container is transshipped during the trip. Note also that each transshipment may involve several loading and unloading operations.

The SRI-axioms we defined in the MCO implicitly describe itinerary patterns resulting in additional costs or in an increased shipping time for the company that performs the shipping, because they carry out operations that appear unnecessary, therefore they can be considered suspicious. Such patterns have been suggested by Customs we are cooperating with [1]. The first SRI-axiom expresses a container itinerary *loop*: the container is loaded on a vessel in a port P; afterward, the vessel reaches another port where it is transshipped on another vessel, that comes back to the port P before reaching the shippment destination. The second pattern defines an *unnecessary transshipment*: the container is transshipped from its original vessel (A) on another vessel (B), in an intermediate port. Comparing the vessels routes, we infer that vessel A and vessel B go to the same destination, therefore the transshipment is unnecessary. However, with this trick, the container may appear to be coming from the starting port of vessel B.

The SRIs formalisation may be extended to other anomalous patterns: it is sufficient to define in the MCO the corresponding logical axioms, and evaluate them to discover the suspicious itineraries. Moreover, this work offers a formal study towards the search of patterns in moving object trajectories, in research areas other than Maritime Security, such as GISScience, for the discovery of traffic patterns and way-finding for urban modelling, and it may be applied to discover anomalies in spatio-temporal sequences formalized as itineraries, for example intrusion detection in secure areas.

4 Conclusions and Future Work

The logic-based formalization of itinerary patterns may pave the way to a methodology for pattern detection in moving object itineraries. A shortcoming of the approach is scalability. To address such problem, as well as incomplete container itineraries, we plan to investigate the use of data mining techniques, combined with logical rules to improve the expressive power of the representation. Another crucial issue we are investigating is the application of time constraints on itineraries for the definition and the evaluation of logical axioms.

References

[1] Alberto V. Donati, Evangelos Kotsakis, Aris Tsois, Francisco Rios, Mauro Zanzi, Aristide Varfis, Thomas Barbas, and Jose Perdigao. Overview of the ConTraffic system. Technical report, JRC, November 2007.

[2] P. Villa and E. Camossi. A description logic approach to discover suspicious itineraries from maritime container trajectories. In C. Claramunt, S. Levashkin, and M. Bertolotto, editors, *GeoS 2011*, volume 6631 of *Lecture Notes in Computer Science*, pages 182–199. Springer Berlin Heidelberg, 2011.

[3] M. Baglioni, J. de Macêdo, C. Renso, R. Trasarti, and M. Wachowicz. Towards semantic interpretation of movement behavior. In *Advances in GIScience*, Lecture Notes in Geoinformation and Cartography, pages 271–288. Springer Berlin Heidelberg, 2009.

[4] V. Bogorny, B. Kuijpers, and L. O. Alvares. ST-DMQL: A semantic trajectory data mining query language. *International Journal of Geographical Information Science*, 23(10):1245–1276, 2009.

Detection of near misses and undesired encounters on the North Sea

Erwin van Iperen
MARIN
Wageningen, The Netherlands
e.v.iperen@marin.nl

Every year, a number of collisions between ships occur at the North Sea. These incidents are reported by the Netherlands Coastguard, who also takes care of the aftermath of the accidents and makes sure that further environmental damage is prevented. Statistics of reported accidents are also fed into SAMSON, a model to quantify risks and safety at sea, developed by MARIN and commissioned by the Ministry of Infrastructure and the Environment.

Accidents are too rare, fortunately, to directly provide information about the current safety levels on particular areas at sea. Valuable information, however, can also be provided by studying so called near miss situations: situations that could have resulted in accidents, but that were narrowly avoided. These situations are, however, often not reported to the coastguard.

The introduction of AIS provides an excellent opportunity to monitor the shipping traffic. The Netherlands Coastguard has placed AIS base stations along the coastline, as well as on platforms at sea, and it can monitor the entire Dutch exclusive economical zone. For MARIN, AIS data provides valuable insights into the behaviour of ships during encounters, which can be used to further improve the SAMSON model. MARIN therefore receives AIS-data from the Netherlands Coastguard to validate and update the information that is used in safety studies with the SAMSON model.

It is, however, impossible for the Coastguard to observe all possible hazardous situations visually. The Ministry of Infrastructure and the Environment, as well as The Netherlands Coastguard, have therefore asked MARIN to develop methods to automatically detect near misses from AIS. This would provide valuable information about the current safety levels at various locations on the North Sea.

The first phase of the project focused solely on a junction in a busy traffic separation scheme, where many ship encounters occur, both crossing encounters as well as overtaking.

For a number of visually observed representative encounters, some normal, some abnormal, various parameters were studied that together showed the anticipating behaviour of ships in an encounter. Parameters were distance, speed, course, closest point of approach (CPA) and time till closest point of approach (TCPA). The CPA during an encounter is the estimated distance at which the ships will pass, if both keep their current speed and course. The first method of detection is a ranking of the encounters based upon the minimal observed CPA value during the period when the TCPA was less then 9 minutes. The choice for 9 minutes was based upon the representative encounters, but the results were found not to be sensitive for this value.

A meeting with members of the SAN (Shipping Advisory board for the North Sea) was held to get advice about the parameters that could indicate a near miss or undesirable situation. Additional to the CPA and TCPA relation, it was concluded that information about the maintained ship domain was most essential to be able to tell wether ships were too close for comfort or not. The ship domain, the area around the own ship that ships wish to keep free of other traffic, depends on a variety of possible factors, such as length, speed, shiptype, current, weather conditions. For the initial study only length was considered as a factor.

An estimate of the maintained shipdomains in the considered area was made by transforming each shiptrack during an encounter to ship coorinates, that is, the relative position of the other ship as seen

from the own ship. This is given by relative bearing and absolute distance. All tracks of encounters can then be plotted over each other, resulting in a cloud of dots and in the middle an more or less empty space, where the own ship is located. The shape of the empty area around the ship is an indication of the maintained domain.

The domain shape was found to be different for the type of encounters: overtaking manoeuvres showed a symmetric ellipse, but crossing encounters showed a bigger ellipse, slightly turned to either right or left, depending on whether ships approached from starboard side or port side. The bigger size of the domain is caused by early anticipation on the domain of the own and other ship, by maintaining for instance a CPA of 0.2 nautical miles during approach. Moreover, ships crossing from starboard side are anticipated upon by the own ship by usually crossing at the stern. In case they pass at the bow, the maintained distance is far bigger.

Given the different domain shapes, the domains were determined for different encounter types: overtaking, crossing from starboard side, crossing from portside. For each type of encounter, the domains around the ships were plotted for 5%, 1% and 0.5% percentiles. This was done for only a month of AIS data. The 0.5% percentiles confirmed some of the opinions of SAN members, that at the front of the ship, the maintained domain would be 1 mile.

If the length of the ship is also taken into account, a second domain estimate was described by the number of shiplengths before and ship breadths at the side of the ship.

Based on the observed domains per encounter type for one month, for all encounters during this month it was determined whether the approaching ship entered the domain. This was done both for the absolute distance domain as the relative-to-length domain. The number of positions in the domain during the encounter, was used to further rank the encounters. This resulted in a number of tugs towing a workship appearing in the top of the rankings. These situations should be left out, also while estimating the domain. For this study, this was not yet done.

In total there are three rankings of encounters: two domain based rankings, and one ranking based on the CPA-TCPA observations.

Out of all the encounters, all the encounters were viewed where ships came within 0.5 mile of each other. Five situations that should surely be found by the detection method, were marked. These were typically encounters where ships heavily and abrubtly anticipated. It was found that neither of the three rankings contained these situations in the top 30 of the ranking. However, after combining the three rankings in an obvious way, all of these five cases were found in the top 35, and three were found in the top 10 of the combined ranking.

The second fase of the near miss detection study will focus on generalisation of the methods, extending it to different areas and over a longer period. The method of ranking will be refined, for example by letting experts judge various situations, such that a clearer training set can be created and the distinction between situations that should and should not be detected, will be improved.

PRESTO:
A Poseidon Research Tool to Create Artificial Vessel Trajectories

Jeroen H.M. Janssens*, Hans Hiemstra⋆, Eric O. Postma*
* Tilburg center for Cognition and Communication, Tilburg University, Warandelaan 2, 5037 AB, Tilburg, The Netherlands, {j.h.m.janssens,e.o.postma}@uvt.nl
⋆ Thales Nederland B.V., P.O. Box 42, 7550 GD, Hengelo, The Netherlands

The automatic detection of anomalies in the maritime domain requires representative anomalous instances. We developed Presto [2], an open-source application (GPLv3 license) that enables experts and researchers in the maritime domain to create artificial anomalous vessel trajectories that may be characteristic of traffic violations, illegal fishing activities, drug smuggling, or piracy. When merged with existing real-world data, the artificial trajectories make possible the usage and evaluation of machine learning algorithms for the automatic detection of anomalies.

1 The need for maritime anomalies

A Maritime Safety and Security (MSS) system aims at guiding a surveillance operator to find maritime anomalies, such as vessel traffic violations, illegal fishing activities, and drug smuggling. In real-world data, serious maritime anomalies rarely occur. Because the MSS system might be deployed in any maritime circumstance, we need to evaluate if our machine learning algorithms [1] can properly detect such anomalies. To this end, we have developed Presto, an application which enables maritime domain-experts to easily create artificial scenario's. In contrast to existing simulation applications, such as VR-Forces [3], which impose restrictive behavior models, our application gives the expert full control over the vessel trajectories.

2 Overview of Presto

The main concept in Presto is a scenario, which can contain one or more trajectories. Each trajectory is defined by several waypoints. A waypoint is a location on the world map and contains the additional parameters velocity, time, and curvature. Figure 1 shows the user interface of Presto, which consists of three main elements: (a) the world map, (b) the timeline, and (c) the property editor. Using the world map, the user can edit the position of the trajectories and their waypoints. The timeline gives an overview of the trajectories within the scenario and is used to navigate through time. The property editor with its three tabs "scenario", "track", and "waypoint", is used to edit element parameters (e.g., for a scenario: name and description; for a trajectory: name and flag; and for a waypoint: velocity and curvature).

Presto has the ability to load existing, real vessel-trajectory data as so-called background data. This background data contains the positions and bearings of all vessels for a certain period (e.g., a day). It serves as a reference when creating new artificial trajectories, so that, for instance, collisions can be avoided or enforced. The generated data is used to create real-world vessel-trajectory data by filtering the appropriate variables such as location, velocity, and bearing, such that data points mimic messages according the Automatic Identification System (AIS) protocol. The exported data is fused with real-world data, such that the artificial data cannot be distinguished up front from the original data.

Presto is entirely programmed in Java, and makes extensive use of the Eclipse Rich Client Platform. The world map is based on NASA World Wind, an open-source alternative to Google Earth. The use of

†This work has been carried out as part of the Poseidon project under the responsibility of the Embedded Systems Institute, The Netherlands. This project is partially supported by the Dutch Ministry of Economic Affairs under the BSIK03021 program.

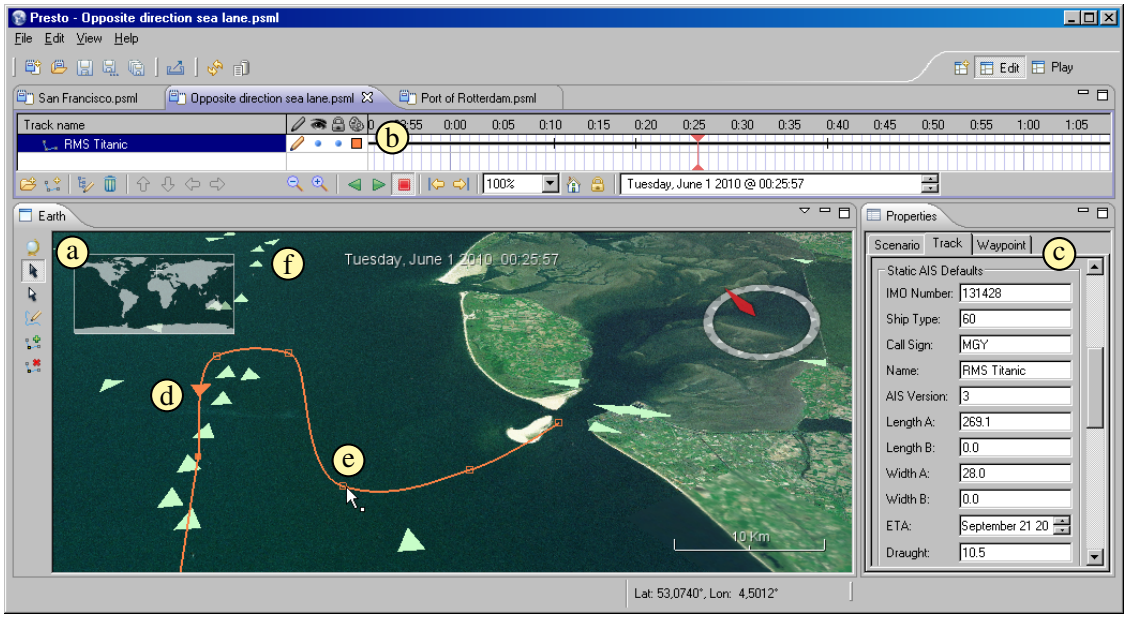


Figure 1: A screenshot illustrating the different user interface elements and concepts of Presto: (a) the world map, (b) the timeline showing the artificial trajectory in the scenario, (c) the property editor showing the trajectory properties, (d) the current location of the artificial vessel along the created trajectory, (e) one waypoint of the trajectory, and (f) the background data at the current moment.

these software solutions makes Presto suitable for recent versions of Windows, Mac OS X, and Linux based operating systems. The world map requires a hardware-accelerated 3D graphics card and an internet connection for downloading the high-resolution terrain images, although these can also be cached.

3 Academic and industrial adoption of Presto

Currently, the beta version of Presto is successfully being used by researchers within the Poseidon project and domain-experts at the Maritime Safety and Security division of Thales Nederland. Additionally, Presto has enabled domain-experts to discuss and exchange potential maritime situations, and to challenge researchers with anomalies that are not present in real-world data. In future work we plan to implement additional import and export functionality such that Presto can be integrated more seamlessly with existing applications that are used within the MSS domain.

References

[1] J.H.M. Janssens, I. Flesch, and E.O. Postma. Outlier detection with one-class classifiers from ML and KDD. In M.A. Wani, M. Kantardzic, V. Palade, L. Kurgan, and Y. Qi, editors, *Proceedings of the 8th International Conference on Machine Learning and Applications*, pages 147–153, Miami, FL, 2009.

[2] J.H.M. Janssens, H. Hiemstra, and E.O. Postma. Creating artificial vessel trajectories with Presto. In *Proceedings of the 22nd Benelux Conference on Artificial Intelligence (BNAIC 2010)*, Luxembourg, Luxembourg, 2010.

[3] MÄK. *VR-Forces: The complete simulation toolkit*, available from: http://www.mak.com/products/vrforces.php. Accessed June 2010.

Author Index