Improving maritime anomaly detection and situation awareness through interactive visualization

Maria Riveiro, Göran Falkman, Tom Ziemke
School of Humanities and Informatics, University of Skövde
541 28 Skövde, Sweden
Email: {maria.riveiro, goran.falkman, tom.ziemke}@his.se

Abstract—**Surveillance of large land, air or sea areas with a multitude of sensor and sensor types typically generates huge amounts of data. Human operators trying to establish individual or collective maritime situation awareness are often overloaded by this information. In order to help them cope with this information overload, we have developed a combined methodology of data visualization, interaction and mining techniques that allows filtering out anomalous vessels, by building a model over normal behavior from which the user can detect deviations. The methodology includes a set of interactive visual representations that support the insertion of the user's knowledge and experience in the creation, validation and continuous update of the normal model. Additionally, this paper presents a software prototype that implements the suggested methodology.**

Keywords: anomaly detection, interaction, visualization, situation awareness, visual data mining, visual analytics.

I. Introduction

Detection of unusual vessel activities has been identified as an important objective for enabling maritime situation awareness in the homeland security domain [1]. Achieving situation awareness is crucial for making effective decisions [2]. However, such awareness in complex situations may be difficult to achieve. Surveillance systems are commonly complex in function and structure because they process and present huge quantities of heterogeneous data from multiple sources (radars, cameras, automatic identification systems, etc.). Monitoring this kind of systems is a challenging activity for humans, due to not only the amount of information, the high number of variables involved or the opacity and complexity of the data mining techniques used in the detection process, but also other factors, like time pressure, high stress, inconsistencies and the imperfect and uncertain nature of the information.

Automatic learning algorithms for anomaly detection are usually based on the assumption that sufficient training data is available and nearly complete with regard to all possible normal behaviors. Otherwise, the learned model of the normality cannot confidently classify new observations as abnormal, since it can just be unseen normal events. This assumption, however, is very optimistic since recorded training data can hardly cover all the possible events that occur in reality. Human expert knowledge can be very valuable in these cases as it can be used to update, refine and improve the normal model and guide the anomaly detection process. However, analytical models for anomaly detection are not necessarily intuitive to humans. Information visualization methods can be of great value when large data sets must be analyzed, but they can also be limited by the problem of data dimensionality, especially when a high number of attributes are to be displayed. We believe that in order to increase the confidence in the detection of anomalies, it is useful to combine the power of computational methods with human background knowledge, experience and flexible thinking using interactive visualization. In this paper, we propose a methodology that makes possible to filter out vessels by building a model of normal behavior from which we can detect deviations. Events that are classified as anomalies can be flagged as alerts for a human operator who continuously validates, updates and refines the normal model by interacting with the visual representations of partial results of the detection process.

The main contribution of this paper is an interactive methodology based on visual representations that involves the user in the anomaly detection process and therefore benefits from the operator's knowledge and experience. Thus, the detection procedure becomes transparent to the user, which increases his/her confidence and trust in the system.

II. Information fusion, HCI and information visualization

In today's information age, the lack of information is seldom a problem. Rather the problem is often the opposite, the overload of information. The difficulty of processing and handling vast amounts of information from many sources is a common feature of many real-life domains. Military operations, crisis management or homeland security applications involve a large number of actors with different characteristics, needs and behaviors. Part of the solution lies in the ability to process and filter the information in a manner that results in knowledge, providing responders and decision makers with improved situation awareness.

Information fusion has been identified as key enabler for providing decision support [3]. It includes theory, techniques and tools for exploiting the synergy in the information acquired from multiple sources, for example sensors observing the environment, databases storing knowledge and simulations predicting future behavior [4].

Most information fusion applications are designed for human decision makers, but perhaps often without sufficient consideration for the actual end users. The lack of research in human-computer interaction related issues has been acknowledged by several researchers in the information fusion community (e.g [5]–[7]). The traditional approach, typically, as illustrated by the JDL model, shows that data flows from sensors (source) toward the human (receiver). This is a somewhat simplistic interpretation though, given that the human is often involved in each step of the fusion process and is not only an information consumer. Using this basic orientation, rich information from multiple sensors is compressed for display on a two-dimensional computer screen (referred to as the "HCI bottleneck" problem by Hall, Hall and Tate [8]).

In order to overcome the HCI bottleneck in the information fusion process and account for functions for information representation and human machine interaction, Hall, Hall and Tate, [5], proposed the introduction of a new level in the JDL model, *level 5: cognitive refinement*. Level 5 accounts for functions to support a human decision-maker in the loop, users in collaborative environments and cognitive aids. Examples of functions for level 5 processing are (adapted from [8]): cognitive aids (functions to aid and assist human understanding and exploitation of data); negative reasoning enhancement (humans have a tendency to seek for information which supports their hypothesis and ignore negative information); uncertainty representation (methods and techniques to improve the representation of uncertainty); focus/defocus of attention (techniques to assist in directing the attention of an analyst to consider different aspects of data) or pattern morphing methods (methods to translate patterns of data into forms that are more easy for an human to interpret). Other authors, like Waltz and Llinas [9], have suggested that the overall effectiveness of a data fusion system is strongly affected by the HCI efficacy.

The use of information fusion techniques can generate vast amounts of complex data that in many cases need to be analyzed by a decision maker. The presentation of the information, the graphical interface and the availability of interaction methods play a central role in the acquisition of the situation awareness necessary to make effective decisions. Advances in data mining, information visualization, interactive computer graphics (software and hardware) and human computer interaction open new possibilities for the access, analysis, navigation and retrieval of information. We believe that methods that support user interaction will bring the best of both sides: human knowledge and experience and the power of automatic processing. Adequate visualization can not only guide the decision making process efficiently but can also support direct interaction with the data, allowing the user to get insights, draw conclusions and, overall, make better decisions.

III. Related work

A. Anomaly detection

Anomalies are defined as deviations from normality. Detecting these deviations can be seen as a classification problem [10]: given a set of observations, they must be classified as normal or abnormal. Nevertheless, conventional classification algorithms cannot be used in real world problems. The main reason is that commonly, only normal samples are available in the training phase, and both normal and abnormal samples are required when conventional classifiers are used. Moreover, the set of anomalies can be infinite since we can encounter unknown anomalies. There are many approaches to anomaly detection in the literature (most of them in the network security arena). The majority of them build a model of the normal behavior in an unsupervised manner. Examples are the work presented in [11] and [12]. Solutions based on artificial immune systems have been also applied in intrusion detection in network systems. An example of the latter is the methodology presented in [10] where fuzzy characterizations of normal/abnormal spaces are used in the detection process.

However, autonomous detection systems are rarely used in the real world [13]. Cain et al. [13] point out that one of the main contributing factors is the difficulty of representing the prior knowledge that the users bring to their tasks. Examples of methodologies for anomaly detection that include human expert knowledge to any extent are rare. An exception is the work presented in [14], where a solution based on Bayesian networks uses the user input to build the normal model in the training phase. Nevertheless, and despite the initial phase, no further input from the user is proposed for updating or validating the results. The methodology presented in this paper tries to overcome this problem, where a set of interactive visual representations support the introduction of the user's knowledge and experience in the creation, validation and continuous update of the normal/abnormal model.

B. Anomaly visualization

Most of the published work regarding anomaly visualization is restricted to the computer security area. Examples of this kind are [15], using 3D displays, [16] regarding network traffic visualization or techniques for visualizing security log-files [17]. Another approach, using Self Organizing Maps (SOM), is the work introduced in [18], where a final representation of the normal/abnormal space is presented using SOM (however, this approach does not include any interaction).

C. Visual data mining

Exploring, analyzing and finding the relevant information in vast amounts of multidimensional sensor data is a complex task. Data mining techniques can filter and extract valuable patterns. The integration of data mining and information visualization techniques has received a lot of attention in recent years [19]. Nevertheless, visualization has been mainly used to provide better understanding of the final results. The need to tightly include the human in the exploration process is now recognized by many authors (e.g. [20]–[24]). *Visual data mining* focuses on integrating the user in the knowledge discovery process using effective and efficient visualization techniques and interaction capabilities. A classification of visual data mining methods regarding data type, visualization technique and the interaction/distortion technique can be found in [20]. Additionally, significant examples of the use of data mining and data visualization can be found in [22].

Several tools/applications that offer a diverse number of data mining and visualization functionalities have been developed to support the various steps of the knowledge discovery process. Nevertheless, their support focuses often on one part of the process, for example: data prepocessing and exploration [25], clustering [26] or classification [27]. An exception is *VidaMine* [28], an overall framework that supports the entire discovery process: planning, data preprocessing, data integration, evaluation and presentation.

An emerging research area in the past years is visual analytics. *Visual analytics* is defined as analytical reasoning supported by highly interactive visual interfaces [29]. Contributions in this area integrate information visualization, interaction and computational analysis in order to transform massive data into knowledge. When the data analyzed is space related (like in our case), models, methods and tools presented in geovisual analytics [30] are worth considering (e.g. [31]).

Information visualization and visual data mining has had little attention in the information fusion community up to date (for exceptions see [32], [33]).

IV. Methodology

Figure 1 shows a schematic diagram of the suggested methodology for detecting anomalies. The "pre-processing data" step includes cleaning, transformation and integration of data functions. After the pre-processing step, visualization methods and automated analysis methods are applied to the data. The "building normal/special behavior model" step creates a model of the normal behavior from observations recorded during a period of time (training data). Once the model is complete, real time observations (test data) are processed in order to detect abnormal behavior. For that, the cumulated probability value of the observed data is calculated and compared to the threshold level. If the probability value is higher than the threshold, the operator is notified (this can be seen as a hypothesis generation process). The user now has to acknowledge the anomaly, modify the model if it is a false alarm or look for more information if he/she cannot make a confident decision. The control of the events is done through a graphical user interface (GUI) that supports interactive visualization at various levels.

We use two layers of visual representations in the anomaly detection process: (1) general views (list of alerts, geographical map, configurable views– different attributes vs time or space, detailed information, etc.) and (2) interactive displays that allow direct manipulation with the data mining module that builds the normal/special behavior model. The set of interactive visual representations support the user in the knowledge discovery process and in the insertion of the user's experience and knowledge in the system.

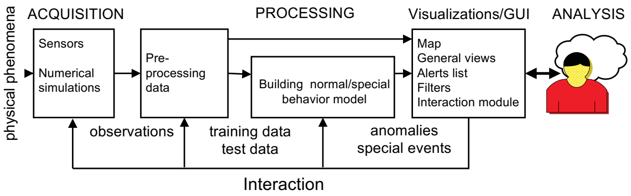


Figure 1. Anomaly detection process. Recorded data is used to build a normal model that is later used to determine the probability a of real time observations from sensors. The interaction capabilities of this methodology support the involvement of the user in every step of the detection process and in the continuous refinement of the normal space characterization.

The purpose of the suggested methodology is to support the acquisition of situation awareness through interactive visualization, from the extraction of the environmental information and its integration with previous knowledge to create a coherent mental picture, to the prediction or anticipation of future events. Furthermore, the interaction capabilities engage the user in the detection process, turning it into a more transparent process that may increase the user's trust in the system.

The following section presents an instantiation of the methodology, section VI presents an overview of the general views and the GUI and, finally, an application example using synthetic data from a region of Sweden is presented in section VII.

V. Applying the methodology to maritime traffic

This section instantiates the usage of the methodology for maritime anomaly detection focusing on the "building the normal/special behavior model" step of the process depicted in figure 1.

A schematic diagram of the approach selected to build the normal/special behavior model is shown in figure 2. This approach is based on the work presented in [12] (a Gaussian Mixture Model (GMM) over a SOM of the training data is used for that). We have extended here their proposal adding an interactive module that allows continuous refinement of the calculated model and development of a "special event" model by the user. This method was selected over other approaches since it generates a visual representation of the normal space and feature clusters that can be used as an interactive basis for refining the model. The graphical representation of the model of the environment facilitates the understanding of the model itself, since peaks and valleys are quickly identified as normal and abnormal behavior (see figure 5).

Essentially, the detector builds a normal behavioral model using a clustering algorithm, the SOM, over the training data set. The SOM learns what is normal via iteration over the training data. In order to quantify probabilities of normal and anomalous behaviors, the detector uses a GMM coupled with the use of Bayes' theorem. Once the model of the normal behavior is established, the detector can be used on the test data (which represent real world observations – sensor readings of vessel movement). For each new observation, $P(observation_{vesselID})$ is calculated. A sliding window over the m most recent observations is used to calculate an average probability value. If the probability value is higher than a given threshold, the detector will flag the vessel as anomalous.

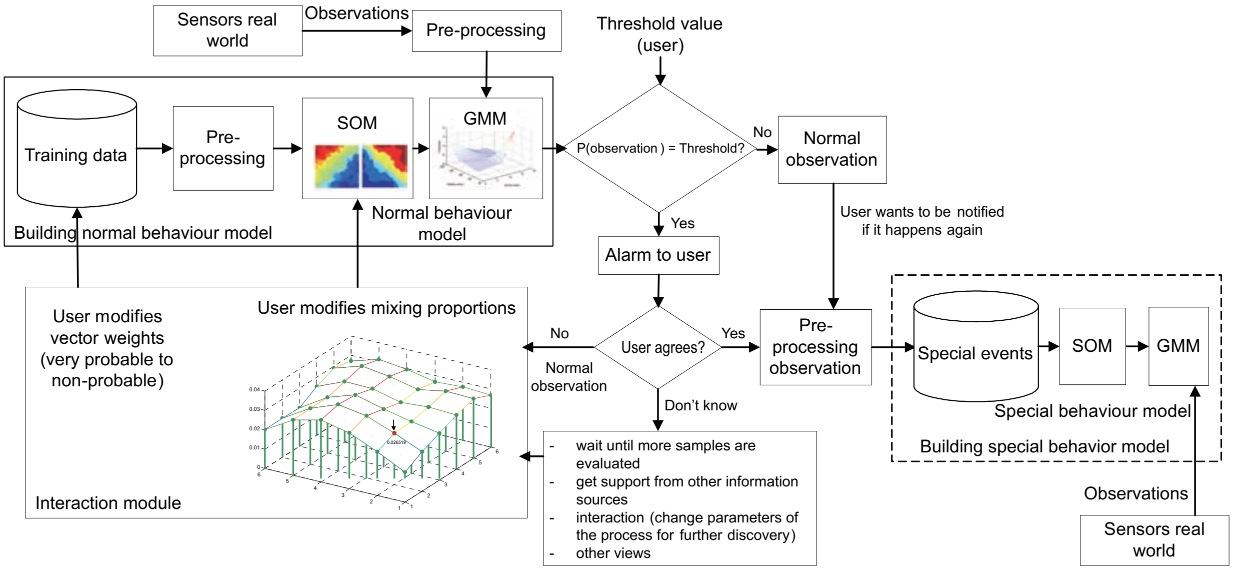

Sensors real world
Observations
Pre-processing
Threshold value (user)
Training data
Pre-processing
SOM
GMM
P(observation) = Threshold?
No
Normal observation
Building normal behaviour model
Normal behaviour model
Yes
Alarm to user
User wants to be notified if it happens again
User modifies vector weights (very probable to non-probable)
User modifies mixing proportions
No
Normal observation
User agrees?
Yes
Pre-processing observation
Special events
SOM
GMM
Don't know
Special behaviour model
Building special behavior model
- wait until more samples are evaluated
- get support from other information sources
- interaction (change parameters of the process for further discovery)
- other views
Observations
Sensors real world
Interaction module

Figure 2. Simplified schema of the anomaly detector (corresponding to the "pre-processing" and "building normal/special behavior model" steps in figure 1). SOM and GMM are used to create a model of the normal behavior over the training data. The user validates this model by interacting with the visual representations of partial results of the algorithm. A model of special events is also created, validated and updated by the user. The operator can set parameters for example, the alarm threshold.

A. Clustering using a Self Organizing Map

A SOM can be seen as a clustering algorithm based on a neural network [34]. It takes a set of n-dimensional training data as input and clusters it into a smaller set of n-dimensional nodes, also known as model vectors. These model vectors tend to move toward regions with a high training data density, and the final nodes are found by minimizing the distance of the training data from the model vectors [35]. The SOM creates a 2D map from n-dimensional input data. In the map, it is usually possible to identify borders that define different clusters [36]. These clusters consist of input data with similar characteristics, in our case, vessels with similar behavior. The output from the SOM is useful for classification, and can be used for portraying a compressed representation of a "normal picture" (see an example in figure 3). However, it does not provide a complete solution to the anomaly detection problem since there are many events that do not clearly fall into these well-defined clusters. Therefore, a GMM has been used on top of the SOM.

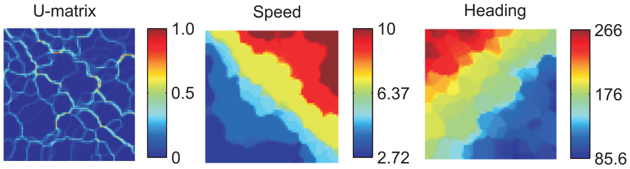


Figure 3. SOM for the HF (cargo ship) vessels. The maps have rectangular structure and the dimensions of the grid are 60 by 60 (D). The first map corresponds to the unified distance matrix (U-matrix) and then the component planes for the features speed and heading. The U-matrix visualizes distances between neighboring map units, and thus shows the cluster structure of the map: high values of the U-matrix indicate a cluster border, uniform areas of low values (blue) indicate the clusters themselves. The component planes show clusters of data with similar values (a color identifies a cluster).

B. Statistical characterization of clusters using a Gaussian Mixture Model

A GMM is a statistical model in which the overall probability distribution, $P(x_1, ..., x_n)$, is synthesized from a weighted sum of individual Gaussian distributions (where the sum always is vaguer than the individual distributions themselves).

$$P(x_1, ..., x_n) = \sum_{i=1}^{D}\sum_{j=1}^{D} p_{ij} P_{ij}(x_1, ..., x_n)\ . \qquad (1)$$

The individual distributions, P_{ij}, in this case correspond to the model vectors that were the output from the SOM. Each model vector is characterized by a n-dimensional Gaussian probability density function. The mean of each individual probability density function is given by the final weights for the model vector, while the variance is given by the dispersion of training data around the model vector. Since the probability density function is a multivariate Gaussian distribution, it can be calculated by

$$P_{ij}(x_1, ..., x_n) = \frac{1}{\sqrt{(2p_{ij})^n}|\Sigma|^{(1/2)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\ , \qquad (2)$$

where μ is the mean vector, Σ is the covariance matrix, and $|\Sigma|$ is the determinant of Σ.

The mixing proportions p_{ij} in equation 1 and 2 are weights of each individual model vector. The mixing proportions correspond to the probability of each map unit to be selected as the best matching unit in the SOM (see figures 4 and 5).

C. Calculation of probabilities using Bayes’ theorem

When new observations arrive, the GMM can be used to quantify the likelihood $P(d|H = normal)$ for obtaining the observation d given the learned model of what is to be considered as a normal event (H is the hypothesis). However, the quantity we want to calculate is the probability of an anomalous event, given the observed data, $P(H = anomalous|d)$ (in order to calculate $P(H = anomalous|d)$, we can just take the complement $1 - P(H = normal|d)$, where $P(H = normal|d)$ is the probability of a normal event given the observed data). To calculate $P(H = normal|d)$ from the likelihood we have to use Bayes’ theorem:

$$P(H = normal|d) = \frac{P(d|H = normal)P(H = normal)}{\sum_{h \in H} P(d|h)P(h)} , \quad (3)$$

where h refers to the hypothesis: being normal or not. The prior probability, $P(H = normal)$ adjusts the detection threshold and can be fine-tuned by the human operator in order to get an acceptable ratio between the detection rate and the false positive rate. Otherwise, it reflects the expected relative frequency between the number of normal observations and the total number of observations. The denominator in equation 3 can be seen as a normalization constant. In order to calculate this normalization constant we need to know the quantity $P(d|H = anomalous)$, which is not known since we have not built such a model. Hence, we conclude:

$$P(H = normal|d) \propto P(d|H = normal)P(H = normal) . \quad (4)$$

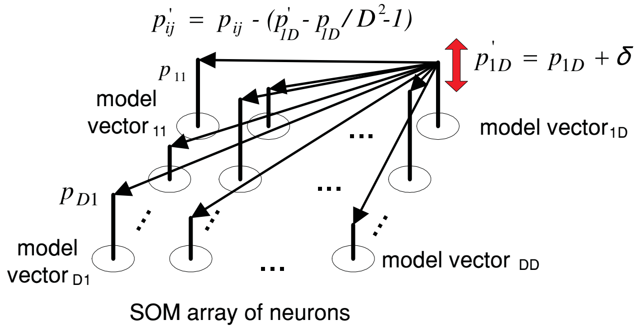


Figure 4. SOM grid: each node represents all the observations that had this neuron as the best matching unit (model vector). The mixing proportions (p_{ij}) represent how probable it is that the neuron is the best matching unit. If the user considers that a value must be modified (p'_{1D} by δ), all the other probability values must be updated ($p'_{ij} = p_{ij} - (\delta/D^2 - 1)$). The normal behavior model is then updated.

D. Interaction module

An average probability value is calculated continuously over the m most recent observations. If the probability value is higher than a given threshold, the detector will flag the vessel as anomalous and alert the operator. If the user considers that it constitutes a real threat and it is representative of abnormal behavior, the observed data will be part of the “special events” database. The “special events” database contains any behavior that the user would like to be alerted of in the future (that includes not only abnormal behavior but also any rare, suspicious or unknown events). However, if the user considers that the flagged vessel exhibits a normal behavior, the normal model must be updated in order to prevent that this false alarm occurs in the future. The user then interacts with the graphical representation of the mixing proportions of the SOM (see figures 4 and 5) or introduce the probability values of the event (from very probable to not probable at all). In the latter, the observations are added to the training data and its weight values are updated regarding the probability value that the operator has introduced. Using coordinated views, the modified values update the main map display and the alert list. Thereby, the model of normal behavior is built and updated by the user.

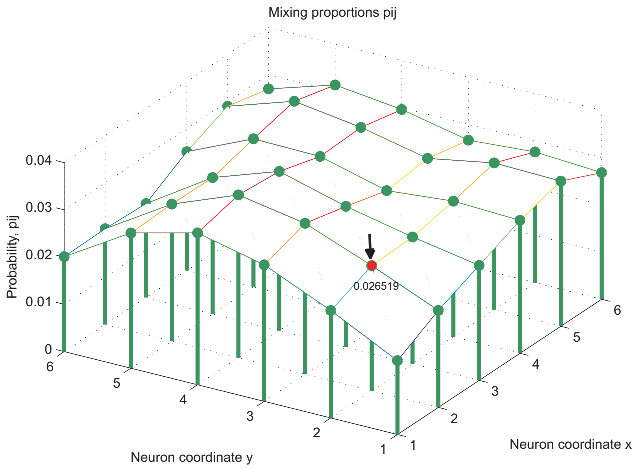


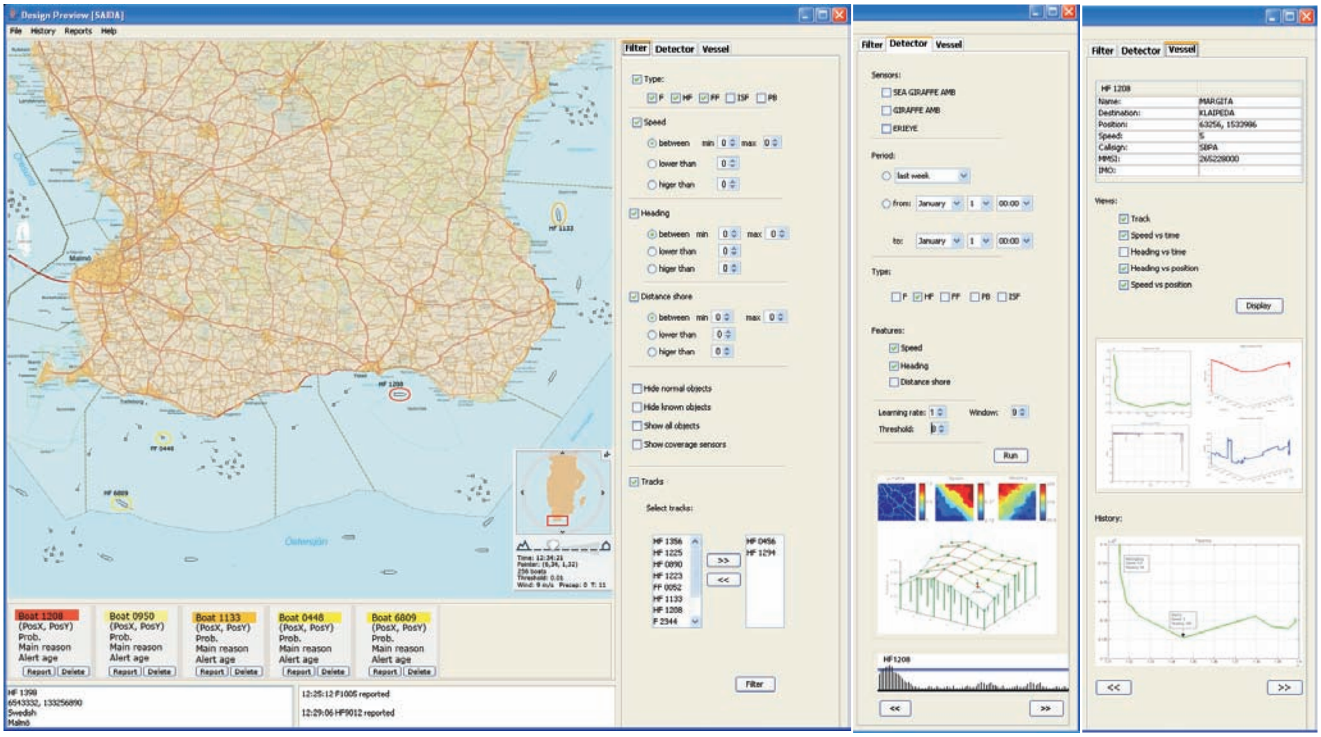
Figure 5. Mixing proportions p_{ij}. Peaks and valleys are quickly identified as normal/abnormal behavior. The user can drag/modify the probability values if he/she considers that they are not realistic. If one probability value is modified, the others are recalculated (see figure 4). This graphical representation provides support for interaction and continuous refinement of the normal model.

VI. Graphical User Interface

The methodology presented in section IV and instantiated in section V has been implemented in a software prototype interface. Its design is based on the suggested requirements as seen in the literature ([37] and [38]) and the needs specified by experts from Saab Microwave Systems (Gothenburg, Sweden). The GUI supports interaction at various levels in order to increase the confidence in the identified anomalies.

The graphical interface has been divided into three main areas or modules (see figure 6): geographical map (left side of the display), controls (right side of the display: filter module, detection module and vessel module) and detailed information and alarms list (bottom left side of the display).

There is a basic task that the operator constantly carries out: establishing and update the normal picture or situation of the supervised zone. Over the normal situation, abnormal behavior must be identified. Therefore, both background awareness



Time: 12:34:21
Pointer: (6,34, 1,32)
256 boats
Threshold: 0.01
Wind: 9 m/s Precap: 0 T: 11

Boat 1208
(PosX, PosY)
Prob.
Main reason
Alert age
Report | Delete

Boat 0950
(PosX, PosY)
Prob.
Main reason
Alert age
Report | Delete

Boat 1133
(PosX, PosY)
Prob.
Main reason
Alert age
Report | Delete

Boat 0448
(PosX, PosY)
Prob.
Main reason
Alert age
Report | Delete

Boat 6809
(PosX, PosY)
Prob.
Main reason
Alert age
Report | Delete

HF 1398
6543332, 133256890
Swedish
Malmö

12:25:12 F1005 reported
12:29:06 HF9012 reported

Filter:

☑ Type:
☑ F ☑ HF ☑ FF □ ISF □ PB

☑ Speed
● between min 0 max 0
○ lower than 0
○ higer than 0

☑ Heading
● between min 0 max 0
○ lower than 0
○ higer than 0

☑ Distance shore
● between min 0 max 0
○ lower than 0
○ higer than 0

□ Hide normal objects
□ Hide known objects
□ Show all objects
□ Show coverage sensors

☑ Tracks
Select tracks:
HF 1356
HF 1225
HF 0890
HF 1223
FF 0052
HF 1133
HF 1208
F 2344

HF 0456
HF 1294

Detector:

Sensors:
□ SEA GIRAFFE AMB
□ GIRAFFE AMB
□ ERIEYE

Period:
○ last week
○ from: January 1 00:00
to: January 1 00:00

Type:
□ F ☑ HF □ FF □ PB □ ISF

Features:
☑ Speed
☑ Heading
□ Distance shore

Learning rate: 1 Window: 9
Threshold: 0

Vessel:

HF 1208	
Name:	MARGITA
Destination:	KLAIPEDA
Position:	63256, 1533986
Speed:	5
Callsign:	SBPA
MMSI:	265228000
IMO:	

Views:
☑ Track
☑ Speed vs time
□ Heading vs time
☑ Heading vs position
☑ Speed vs position

Figure 6. Graphical user interface: filter, detector and vessel module.

(keeping the "big picture") and foreground awareness (particular details) must be supported. The vessels are displayed over a background map of the area. There is an overview map that shows in which area the cursor is at the moment (this map facilitates navigation, zooming in and out and displays information of the area: exact coordinates of the cursor, temperature, precipitations, wind, etc.). Different icons are used to represent vessel type. Part of their tracks are displayed (positions reported by the sensors in the last hour). Longer tracks indicate high speed. If the vessel is flagged as anomalous, its identification number and a colored ellipse will appear around the icon. The color reflects the probability of the anomaly (red, orange, yellow).

Below the main map, detailed information on the alerts, selected vessel and reported alerts is displayed. When a vessel is flagged as anomalous, a card with its information appears in the alert list (with information such as object ID, exact coordinates, probability of being anomalous, age of the alert, delete and report buttons, etc.). The right side of the display shows the reported alerts to the control center.

Filtering options are grouped under the filter tab. The operator can filter the displayed vessels regarding: vessel type, attribute values, display complete tracks for the selected objects, show all the vessels or detected objects, show coverage from our sensors, etc.

The next tab groups functions related with the detection process. The operator can select the source of the data since different sensors have different coverage, resolution and error margins. The period of time of the training data can be also chosen. Additionally, the operator can select the type of vessel, the features involved in the detection process, the values of the learning rate (SOM), sliding window number of samples (m) and threshold value. Under this tab the operator can control the "building the normal/special behavior model" module, e.g. manipulating the mixing proportions display (figure 5). Cumulative probability values and the threshold value are displayed over time (the operator can adjust the threshold value in real time, thus controlling the false alarm rate). Views of the continual evolution of the model vectors for a given boat can be found under the tab "vessel". Here, the history of the vessel can be displayed. Parallel views and 3D views of individual features vs. position are available.

VII. Application Example

In order to test the approach presented here and exemplify the visual interactive components, we have used training and test data provided by Saab Microwave Systems. The data has been generated using the ground target simulator GTSIM, described in [39]. Since the training data is synthetic, it does not include any anomalies. The original data consists of a large set of observations with a number of attributes: time stamp, object-ID (object identification number), object type (classification, e.g. fishing boat, cargo, etc.) and position (given by x-, y- and z-coordinates using the Swedish grid, RT90). There are some measurement errors associated with the objects positions, otherwise there are no uncertainties in the observations.

The training data set has been preprocessed by cutting out a region of interest, in this case a region south of Sweden. The original attributes have been preprocessed into the following attributes: time stamp, vessel type, x-coordinate, y-coordinate, z-coordinate, speed and heading. There are three types of vessels: F ('Fartyg', vessel), FF ('Fiskefartyg', fishing boat) and HF ('Handelsfartyg', cargo ship).

The initial experiments have focused on the cargo ship vessels, HF, since their behavior is easier to analyze than, for example, the fishing boats. For testing our approach, anomalies were hidden in the synthetic test data. Two different types of anomalies were found using the methodology described in previous sections: (1) one of the vessels has abnormal speed values, compared to the training data and (2) a HF vessel approaches the coast (this constitutes an abnormality since this behavior has not been seen before in the analyzed area, here, the heading values are not considered normal). Nevertheless, in order to detect the vessel approaching the coast line, the threshold value must be reduced considerably, which generates high number of false positives. There are also other anomalies in the test set which our detector has not been able to find, for example HF vessels acting as fishing boats. We believe that the number of observations m used to calculate the average probability value must be higher in these cases, since the abnormality occurs in many samples that represent the track–positions. But a high m value will mask abnormal behavior that occurs in few samples. Thus, the selection of m is of high importance since will affect the detection rate and the false positive rate. Another solution is to calculate different average values for different m values.

In conclusion, preliminary results show that this approach produces satisfactory outcomes, since single attribute anomalies can be detected. However, the experiments carried out evaluate the performance of the anomaly detector, but no user tests have been performed so far to really prove the effectiveness or usefulness of the complete methodology. User evaluations will be carried out in the near future.

VIII. Conclusions and future work

Surveillance applications are a clear example where vast amounts of multidimensional sensor data from a large number of objects with different characteristics and behaviors are processed. Finding relevant patterns and special events in them is normally a difficult task that can rarely be solved in a fully automatic manner. We believe that data mining methods that support user interaction integrate the best of both sides, human knowledge and the power of automatic processing. Nevertheless, human interaction is not easy to include in complex cases. Visual representation of partial results of the automatic process and interactive views can support the involvement of the user in the detection process, bringing his/her knowledge and experience into the system. This paper has presented and exemplified a methodology that combines human expert knowledge and data mining techniques through interactive visualization in order to discover anomalies in maritime traffic. Although the methodology described in section IV has been developed for maritime anomaly detection in particular, it has much wider application in general (e.g. it could also be used in network intrusion detection).

Future improvements to this method will involve the incorporation of relational information (relations with other vessels or objects in the environment). The evaluation of this methodology will be complemented with tests on real world data. In this case, modifications or extensions to the methodology described here can be required, since the assumption of no abnormal behavior in the training data may not hold. Furthermore, other important issues, like how to store the information in the normal/special behavior databases, will be considered.

In the future, studies will be undertaken to further evaluate the usefulness of the suggested methodology. However, objective measures of effectiveness for visual data mining techniques are rare. Methods used in the HCI community can be applied here (for example, the usefulness of a software product can be defined by its usability and utility).

Acknowledgements

This work was supported by the Information Fusion Research Program (University of Skövde, Sweden) in partnership with the Swedish Knowledge Foundation under grant 2003/0104 and carried out in collaboration with Saab Microwave Systems (Gothenburg, Sweden). We would like to thank Thomas Kronhamn, Martin Smedberg and Håkan Warston (Saab Microwave Systems) for providing the data and for their valuable feedback. Thanks also to the other members of the ground situation awareness scenario, for their suggestions and fruitful discussions.

The SOM calculations were partially implemented using the Matlab SOM toolbox provided by Helsinki University of Technology, Finland.

References

[1] N. Bomberger, B. Rhodes, M. Seibert, and A. Waxman, "Associative learning of vessel motion patterns for maritime situation awareness," in *Proceedings of the 9th International Conference on Information Fusion*, July 2006.

[2] M. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors Journal*, vol. 37(1), pp. 32–64, 1995.

[3] J. Roy, R. Breton, and S. Paradis, "Human-computer interface for the study of information fusion concepts in situation analysis and command decision support systems," in *Signal Processing, Sensor Fusion, and Target Recognition X*, I. Kadar, Ed., vol. 4380. SPIE, 2001, pp. 361–372.

[4] B. Dasarathy, "Information fusion – what, where, why, when, and how?" *Information Fusion*, vol. 2, pp. 75–76, 2001.

[5] M. J. Hall, S. A. Hall, and T. Tate, "Removing the HCI Bottleneck: How the Human Computer Interaction (HCI) affects the performance of Data Fusion System," in *2000 MSS National Sysmposium on Sensor and Data Fusion*, 2000, pp. 89–104.

[6] E. Blasch and S. Plano, "JDL level 5 fusion model: user refinement issues and applications in group tracking," in *Proc. SPIE, Signal Processing, Sensor Fusion, and Target Recognition XI*, I. Kadar, Ed., Jul. 2002, pp. 270–279.

[7] D. Hall and S. A. H. McMullen, Eds., *Mathematical Techniques in Multisensor Data Fusion*, 2nd ed. Norwood, MA: Artech House, Inc., 2004.

[8] M. Hall, S. Hall, and T. Tate, *Removing the HCI Bottleneck: How the Human-Computer Interface (HCI) Affects the Performance of Data Fusion Systems*. Boca Raton: CRC Press, 2000, ch. 19 in Handbook of multisensor data fusion .

[9] E. Waltz and J. Llinas, Eds., *Multisensor Data Fusion*. Norwood, MA: Artech House, Inc., 1990.

[10] F. González, D. Dasgupta, and R. Kozma, "Combining negative selection and classification techniques for anomaly detection," in *Congress on Evolutionary Computation*. IEEE, 2002, pp. 705–710.

[11] D. E. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. 13, no. 2, pp. 222–232, 1987.

[12] J. Kraiman, S. Arouh, and M. Webb, "Automated anomaly detection processor," in *Proceedings of SPIE: Enabling Technologies for Simulation Science VI*, A. Sisti and D. Trevisani, Eds., Jul 2002, pp. 128–137.

[13] Y. Cai, R. Stumpf, T. Wynne, M. Tomlinson, S. H. Chung, X. Boutonnier, M. Ihmig, R. Franco, and N. Bauernfeind, "Visual transformation for interactive spatio-temporal data mining," *Journal of Knowledge and Information Systems*, vol. 11, no. 5, pp. 119–142, 2007.

[14] F. Johansson and G. Falkman, "Detection of vessel anomlies - a Bayesian network approach," in *Proceedings of the 3rd International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2007.

[15] S. Musa and D. Parish, "Visualising communication network security attacks," in *IV '07: Proceedings of the 11th International Conference Information Visualization*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 17–22.

[16] I. V. Onut and A. A. Ghorbani, "A novel visualization technique for network anomaly detection," in *Proceedings of the Second Annual Conference on Privacy, Security and Trust*, 2004.

[17] S. T. Teoh, K. L. Ma, S. F. Wu, and T. J. Jankun-Kelly, "Detecting flaws and intruders with visual data analysis," *IEEE Computer Graphics and Applications*, vol. 24, no. 5, pp. 27–35, 2004.

[18] F. A. González, J. C. Galeano, D. A. Rojas, and A. Veloza-Suan, "Discriminating and visualizing anomalies using negative selection and self-organizing maps," in *GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2005, pp. 297–304.

[19] G. Manco, C. Pizzuti, and D. Talia, "Eureka!: an interactive and visual knowledge discovery tool," *Journal of Visual Languages & Computing*, vol. 15, no. 1, pp. 1–35, 2004.

[20] D. A. Keim, "Information visualization and visual data mining," *IEEE Transactions on Visualization and Computer Graphics*, vol. 7, no. 1, pp. 1–8, 2002.

[21] P. C. Wong, "Visual data mining," *Computer Graphics and Applications, IEEE*, vol. 19, no. 5, pp. 20–21, 1999.

[22] U. Fayyad, G. G. Grinstein, and A. Wierse, Eds., *Information visualization in data mining and knowledge discovery*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002.

[23] M. F. de Oliveira and H. Levkowitz, "From visual data exploration to visual data mining: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 3, pp. 378–394, 2003.

[24] R. Burkhard, G. Andrienko, N. Andrienko, J. Dykes, A. Koutamanis, W. Kienreich, R. Phaal, A. Blackwell, M. Eppler, J. Huang, M. Meagher, A. Grün, S. Lang, D. Perrin, W. Weber, A. Moere, B. Herr, K. Börner, J. Fekete, and D. Brodbeck, "Visualization summit 2007: ten research goals for 2010," *Information Visualization*, vol. 6, pp. 169–188, 2007.

[25] L. Yang, "Interactive exploration of very large relational datasets through 3D dynamic projections," in *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining*. New York, NY, USA: ACM, 2000, pp. 236–243.

[26] C. C. Aggarwal, "A human-computer cooperative system for effective high dimensional clustering," in *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*. New York, NY, USA: ACM, 2001, pp. 221–226.

[27] M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel, "Visual classification: an interactive approach to decision tree construction," in *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*. New York, NY, USA: ACM, 1999, pp. 392–396.

[28] S. Kimani, S. Lodi, T. Catarci, G. Santucci, and C. Sartori, "VidaMine: a visual data mining environment," *Journal of Visual Languages & Computing*, vol. 15, no. 1, pp. 37–67, 2004.

[29] J. Thomas and K. Cook, Eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alametos, CA: IEEE Computer Society, 2005. [Online]. Available: http://nvac.pnl.gov/agenda.stm

[30] G. Andrienko, N. Andrienko, P. Jankowski, D. Keim, M. Kraak, A. MacEachren, and S. Wrobel, "Geovisual analytics for spatial decision support setting the research agenda," *Journal of Geographical Information Science*, vol. 21, no. 8, pp. 839–857, 2007.

[31] D. Guo, "Visual analytics of spatial interaction patterns for pandemic decision support," *International Journal of Geographical Information Science*, vol. 21, no. 8, pp. 859–877, 2007.

[32] J. Blythe, M. Patwardhan, T. Oates, M. desJardins, and P. Rheingans, "Visualization support for fusing relational, spatio-temporal data: Building career histories," in *Information Fusion, 2006 9th International Conference on*, Florence, Italy, 2006.

[33] M. Mandiak, P. Shah, Y. Kim, and T. Kesavadas, "Development of an Integrated GUI Framework for Post-Disaster Data Fusion Visualization," in *Information Fusion, 2005 8th International Conference on*, vol. 2, Philadelphia, PA, USA, 2005.

[34] T. Kohonen, "The Self-Organizing Map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[35] J. B. Kraiman, S. L. Arouh, and M. L. Webb, "Automated anomaly detection processor," in *Proceedings of SPIE: Enabling Technologies for Simulation Science VI*, A. F. Sisti and D. A. Trevisani, Eds., Jul 2002, pp. 128–137.

[36] T. Kohonen, *Self-Organizing Maps*, 2nd ed., ser. Springer series in information sciences. Berlin: Springer, 1997.

[37] D. Gouin, P. Evdokiou, and R. Vernik, "A showcase of visualization approaches for military decision makers," in *RTO IST Workshop on Massive Military Data Fusion and Visualisation: Users Talk with Developers, RTO-MP-105*, Halden, Norway, 2002.

[38] M. Barnes, "The human dimension of battlespace visualization: Research and design issues," Army Research Laboratory, Tech. Rep. ARL-TR-2885, February 2003.

[39] H. Warston and H. Persson, "Ground surveillance and fusion of ground target sensor data in a network based defense," in *Proceedings of the 7th International Conference on Information Fusion*, 2004, pp. 1195–1201.