Proceedings of the ASME 2019 38th International Conference on Ocean, Offshore and Arctic Engineering
OMAE 2019
June 9-14, 2019, Glasgow, Scotland

OMAE2019-95963

AN AIS-BASED MULTIPLE TRAJECTORY PREDICTION APPROACH FOR COLLISION AVOIDANCE IN FUTURE VESSELS

Brian Murray*
Department of Technology and Safety
UiT - The Arctic University of Norway
9037 Tromsoe, Norway
Email: brian.j.murray@uit.no

Lokukaluge P. Perera
Department of Technology and Safety
UiT - The Arctic University of Norway
9037 Tromsoe, Norway
Email: prasad.perera@uit.no

ABSTRACT

This study presents a method to predict the future trajectory of a target vessel using historical AIS data. The purpose of such a prediction is to aid in collision avoidance in future vessels. The method presented in this study extracts all trajectories present in an initial cluster centered about a vessel position. Features for each trajectory are then generated using Principle Component Analysis and used in clustering via unsupervised Gaussian mixture modeling. Each resultant cluster represents a possible future route the vessel may follow. A trajectory prediction is then conducted with respect to each cluster of trajectories discovered. This results in a prediction of multiple possible trajectories. The results indicate that the algorithm is effective in clustering the trajectories, where at least one cluster corresponds to the true trajectory of the vessel. The resultant predicted trajectories are also found to be reasonably accurate.

NOMENCLATURE

c	Class
C	Cluster
COG	Course over Ground [°]
e	Eigenvector
E	Eigenvector Matrix
I	Identity Matrix
L	Number of Data Points in Selected Trajectory
M	Number of Models in Gaussian Mixture Model
N	Number of Trajectories
r_s	Search Radius [NM]
S	Vessel State
SOG	Speed over Ground [kn]
t	Trajectory Feature Vector
T	Elapsed Time [s]
T_{pred}	Desired Prediction Time Horizon [s]
x	Reduced Feature Vector
Z	Arbitrary Matrix
ΔL	Step Size [NM]
ϕ	Latitude [DD]
λ	Longitude [DD]
Λ	Eigenvalue Matrix
μ	Mean Vector
π	Prior Distribution
Σ	Covariance Matrix
θ	Rotation Angle [°]
Θ	Model Parameters
Subscripts	
0	Initial State
k	k^{th} State
l	Number of Eigenvectors
m	Model Number in Gaussian Mixture
δ	Maximum Offset
Superscripts	
$\hat{}$	Estimated Parameter / State

*Address all correspondence to this author.

Acronyms

AIS	Automatic Identification System
EM	Expectation Maximization
GMM	Gaussian Mixture Model
PCA	Principle Component Analysis

INTRODUCTION

Maritime situational awareness [1] is vital to ensure safe ship operations. With the advent of autonomous ships, this aspect of ship operations will only increase in importance. One of the major challenges facing autonomous ships is conducting effective collision avoidance maneuvers. As such, the risk of collision must be evaluated continuously, and counter-measures implemented in an optimal manner [2]. Such measures will rely on the situational awareness of the autonomous ship. This entails the ability to predict its future states, as well as the future states of target vessels.

A solution to the collision avoidance challenge is to continue on a predefined course until a collision appears imminent based on the relative velocity of a target vessel and determining the closest point of approach (CPA) between it and the own ship [3]. This implies a close encounter situation, and a reactive collision avoidance maneuver must be conducted. Work has also been done to develop more advanced algorithms, i.e. Kalman filter and extended Kalman filters, to estimate such ship trajectories and subsequently utilize them for collision avoidance maneuvers [4].

An alternative approach involves predictive collision avoidance. In this approach, the trajectory of the own ship and target vessels are predicted, and the future risk of collision evaluated for a given prediction horizon. Such predictions allow for simple collision avoidance measures to be conducted far in advance such as minor speed or course alterations. If effective, such predictive collision avoidance will prevent close encounter situations from occurring at all, enhancing the overall safety and efficiency of the operation.

Predicting ship behavior far in advance is, however, not straight forward, as the future intentions of the target vessels are not known. Historical AIS data contain a wealth of information related to ship behavior for specific regions. By intelligently exploiting the data available, the future trajectory of a vessel can be predicted based on the past trajectories of similar vessels in the same region. [5] provides a comprehensive survey of intelligent maritime navigation techniques utilizing AIS data.

In this paper, historical AIS data are exploited to predict the future trajectory of a target vessel for prediction horizons up to 30 minutes into the future. A subset of data is extracted from the AIS data set based on an initial clustering centered about the target vessel state. This subset of data is comprised of relevant trajectories with respect to the target vessel. These trajectories are then clustered using unsupervised Gaussian mixture modeling. In order to effectively cluster the trajectories, a subset of parameters are generated via Principle Component Analysis for each trajectory and used as input to the clustering algorithm. The resultant clusters represent multiple possible routes the target vessel may follow. The clustering based iterative prediction algorithm utilized in [6] is then applied to the data in each route cluster resulting in a prediction of multiple possible target vessel trajectories. The accuracy of the prediction algorithm is then evaluated by testing 100 random vessel states.

RELATED WORK

AIS data has been the subject of an increasing amount of research in recent years. Most work has however focused on long term time horizons in relation to predicting vessel trajectory patterns and general traffic behavior e.g. [7–10]. Other work includes neural network approaches to vessel trajectory prediction aimed towards aiding vessel traffic management systems [11]. [12] presents a Bayesian network based vessel position prediction algorithm with a particle filter used for prediction horizons in the order of hours. There has been more limited work on short term predictions (order 5-30 minutes). [13] and [14] present AIS-based approaches to predict short term vessel trajectories.

Other work on vessel trajectory clustering techniques include [15] where Dynamic Time Warping and Principle Component Analysis are utilized to generate features for trajectory clustering. Additionally, [16] utilizes Gaussian mixture modeling to evaluate anomalous ship behavior on a sub-trajectory scale.

METHODOLOGY

In this section, the methodology of the vessel trajectory prediction is covered. The method is four-fold. The first step of the method covers the selection of relevant data. The result of this is the extraction of relevant vessel trajectories from historical AIS data with respect to the selected vessel. In the second step, features are generated based on the extracted trajectories for the purpose of enhanced class discriminatory properties for use in clustering. In the third step, the extracted trajectories are clustered, where the clusters represent possible routes that the selected vessel may follow. In the fourth step, a prediction algorithm is applied to each unique cluster resulting in a prediction of multiple possible trajectories for the selected vessel.

Data Selection

The initial vessel state S_0 of the selected vessel is defined in Equation (1):

$$S_0 \rightarrow [\phi_0, \lambda_0, COG_0, SOG_0, T_0] \tag{1}$$

The parameters are measured by the on-board sensors of the own

ship. This initial state will define the basis for the selection of relevant data for the prediction algorithm.

Initial Clustering. It is desirable to identify historical ship trajectories that have a high degree of similarity to S_0. The idea is that if a ship was in the vicinity of S_0 and had similar SOG and COG values, the selected vessel will likely have a similar trajectory in the future. As such, an initial cluster is created to identify similar trajectories.

Consider the matrix Z of spatial data contained in the AIS data set. A rotational affine transformation from one coordinate system to another can be defined such that $Z = [x_z, y_z]$ is rotated by $\theta = COG_0$. The new matrix $Z' = [x_{z'}, y_{z'}]$ is defined as:

$$Z' = RZ^T \tag{2}$$

Where $x_z \in \mathbb{R}$, $y_z \in \mathbb{R}$, $x_{z'} \in \mathbb{R}$, $y_{z'} \in \mathbb{R}$ and R is the rotation matrix defined as:

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{3}$$

Z' will have basis vectors comprised of a vector in the direction of COG_0 and a vector orthogonal to COG_0. An initial cluster C_0 is then created in the Z' vector space, where the distance to all vectors in the direction orthogonal to COG_0 is less than 1 NM, and in the direction of COG_0 less than 0.25 NM. This results in a rectangular cluster in the original vector space as illustrated in Figure 1.

Subsequently, the clustering is extended such that all data points fulfill Equations (4) and (5). This results in an initial cluster about SOG_0 that contains data relating to vessels that had a similar course and speed and were at a similar position along the prevailing route in the region. Such a rectangular cluster orthogonal to COG_0 should capture most vessels that have similar characteristics in the AIS data set.

$$|COG_i - COG_0| \leq COG_\delta \, \forall i \in C_0 \tag{4}$$

$$|SOG_i - SOG_0| \leq SOG_\delta \, \forall i \in C_0 \tag{5}$$

Trajectory Extraction. Using the initial clustering technique outlined, the method continues by identifying unique vessel trajectories within the initial cluster. Once unique instances of vessel trajectories are identified, the earliest instance of each is labeled as the initial point of the trajectory. The trajectory is then extracted from this point and a period of time into the future corresponding to the desired prediction horizon, T_{pred}, with an additional 5 minutes added to allow for sufficient data density at the culmination of a trajectory prediction. This trajectory extraction technique is similar to the Multiple Trajectory Extraction Method outlined in [6] but with an alternate initial clustering.

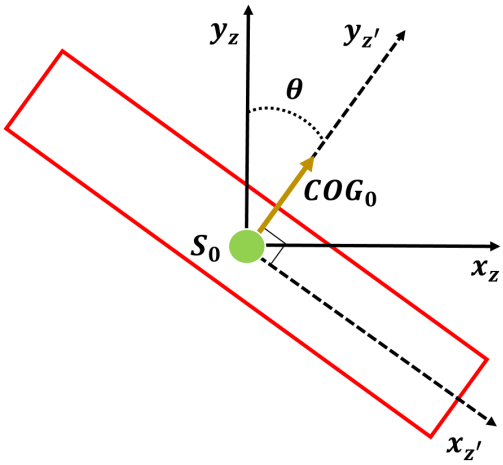

FIGURE 1. ILLUSTRATION OF CLUSTERING VIA COORDINATE SYSTEM TRANSFORMATION.

All extracted trajectories are linearly interpolated at 30 second intervals from the starting point to generate higher density data, as well as provide a common time index utilized for feature generation.

Feature Generation

In this section, the method utilized to generate features for each unique extracted trajectory is outlined. The term feature refers to an individual measurable parameter that describes the trajectory. These trajectory features are used to cluster the trajectories into separate classes. Therefore, it is desirable to generate features that give a high degree of discrimination between trajectories such that the underlying classes can be discovered. This method of clustering and classification is also known as unsupervised learning.

Trajectory Feature Vector. Each unique trajectory extracted from the initial cluster C_0 is stored as a matrix of size $LX4$. Each column contains the parameter values of ϕ, λ, COG or SOG. Each row of the data cluster represents a time instant at 30 second intervals of vessel states along the respective trajectory. Given that the data was interpolated at 30 second intervals, the parameter values describing the trajectories can be directly compared at the same time instances defined, where $T = 0$ is de-

fined at the initial data point. Therefore, a feature vector for each unique trajectory is created by concatenating the columns of the matrix to result in a feature vector $t \in \mathbb{R}^{4L\times 1}$ which is utilized for further analysis.

Dimensionality Reduction. The goal of feature generation is to generate new features that can yield a higher degree of discrimination between classes. In this case, each class represents a cluster of trajectories or traffic route. The aim is to classify each trajectory in an unsupervised manner. As such, the discriminatory properties of the feature vector for each trajectory must be maximized.

If one were to utilize the feature vector previously outlined, one would have a very high number of features in each vector. Take the case of a 30 minute prediction. This would yield a feature vector of length 280. Using so many features will not be an effective manner of conducting unsupervised classification as many features will not yield discriminatory properties. For this case, such features will result from similar trajectory properties between classes, such as when ships sail through a constrained waterway at similar speeds. Using the data from this sub-trajectory will not aid identifying unique trajectories as they will yield very little discrimination between, and will in fact contribute towards clustering the trajectories as part of the same class.

One method to ameliorate this is the the Karhunen-Loéve transform [17], also known as Principal Component Analysis (PCA) [18, 19]. The aim of the transform is to gain uncorrelated features. This is done via the transform shown in Equation (6). Matrix E consists of the eigenvectors of covariance matrix Σ of the set all feature vectors. Λ is the eigenvalue matrix. The relationship is shown in Equation (7). Equation (6) projects a feature vector t onto the space spanned by the eigenvectors. However, many of the eigenvectors that span the space describe very little of the variation in the data.

$$x = E^T t \tag{6}$$

Where $x \in \mathbb{R}^{4L\times 1}$, $t \in \mathbb{R}^{4L\times 1}$ and $E \in \mathbb{R}^{4L\times 4L}$

$$\Sigma = E\Lambda E^T \tag{7}$$

Where $\Sigma \in \mathbb{R}^{4L\times 4L}$ and $\Lambda \in \mathbb{R}^{4L\times 4L}$

In the PCA method, one projects t onto a subspace spanned by the l largest eigenvectors, and as such reduces the number of dimensions in the feature space. As such, the projection of each feature vector onto the subspace spanned by the l largest eigenvectors will result in a new feature vector of length l that describes the most variation in the data. The number of eigenvectors chosen should explain at least 95 % of the variance in the data. This is evaluated by investigating the eigenvalues of the chosen eigenvectors and their sum over the sum of all eigenvalues [20].

In this study, the eigenvectors corresponding to the 3 largest eigenvalues are chosen, i.e. $l = 3$ due to the degree to which the variation in the data was described. It is also possible to view the relationship in the data more easily with 3 eigenvectors. The new reduced feature vector generated by PCA is denoted x, and is generated in Equation 8. This is conducted for all extracted trajectories.

$$x = E_l^T t \tag{8}$$

Where $x \in \mathbb{R}^{l\times 1}$ and $E_l \in \mathbb{R}^{l\times l}$

Trajectory Clustering

Given the trajectory features generated via PCA, the trajectories are clustered into an unspecified number of classes. The number is unspecified as there can be any number of traffic routes present in the data based on S_0. Therefore, the algorithm must discover the most likely number. This is achieved by unsupervised Guassian Mixture Modeling (GMM) through applying the Expectation Maximization (EM) algorithm.

Unsupervised Gaussian Mixture Modeling. The fundamental idea behind Gaussian Mixture Modeling [21] is that a set X of data points is comprised of a mixture of M different Gaussian distributions, each with a mean μ_m, covariance matrix Σ_m and prior distribution π_m. Additionally, a class membership parameter z_i is introduced for each data point x_i. z_i is a vector of length M where $z_{im} = 1$ if x_i belongs to class m and $z_{ik} = 0 \forall k \neq m$. This gives a class conditional probability shown in Equation (9). The most likely model is discovered by maximizing the log-likelihood.

$$p(x_i|z_{im} = 1) \sim N(\mu_m, \Sigma_m) \tag{9}$$

The EM algorithm begins by initializing all parameters. One approach is to initialize all μ_m as random data points x_i, $\pi_m = \frac{1}{N}$ and $\Sigma_m = I$. Using this initialization, the expectation step of the algorithm evaluates the expected class membership $\langle z_{im} \rangle$ shown in Equation (10).

$$\langle z_{im} \rangle = \frac{p(x_i|z_{im}=1;\Theta)\pi_m}{\sum_{k=1}^{M} p(x_i|z_{ik}=1;\Theta)\pi_k} \tag{10}$$

Once this step is done, the data points will have class memberships based on the M initialized Gaussian distributions. Given these memberships, the maximization step takes the log-likelihood and maximizes it with respect to the parameters, Θ. The estimated parameters are given in Equations (11), (12) and (13).

$$\hat{\mu}_m = \frac{\sum_{i=1}^{N} \langle z_{im} \rangle x_i}{\sum_{i=1}^{N} \langle z_{im} \rangle} \tag{11}$$

$$\hat{\Sigma}_m = \frac{\sum_{i=1}^{N} \langle z_{im} \rangle (x_i - \mu_m)(x_i - \mu_m)^T}{\sum_{i=1}^{N} \langle z_{im} \rangle} \tag{12}$$

$$\hat{\pi}_m = \frac{\sum_{i=1}^{N} \langle z_{im} \rangle}{N} \tag{13}$$

This process is then repeated where the expectation step is now based on the updated model parameters. The expectation and maximization steps are repeated until there is little to no change in the model parameters or the total log-likelihood converges.

Upon convergence, the GMM will consist of M different Gaussian distributions that describe the class conditional probabilities of the underlying data. A data point x can then be classified using Bayesian classification, where the class c_m will be chosen such that:

$$p(c_m|x) > p(c_i|x) \, \forall i \neq m, i \in M \tag{14}$$

The posterior probability of class c_m is evaluated from the class conditional probabilities and priors discovered by the EM algorithm via Bayes Rule in Equation (15):

$$p(c_m|x) = \frac{p(x|c_m)\pi_m}{p(x)} \tag{15}$$

This classification is applied to all $x \in X$ resulting in a clustering or classification of the data to M different classes.

Model Selection. The GMM approach requires that number of underlying models, M is provided as input to the algorithm. It is assumed that the number of data clusters or classes is unknown, and can vary from case to case depending on the region in which the trajectory prediction is required. Hence, a flexible model that can adapt to the parameter distribution of the data set is needed.

One method to discover the most likely number of underlying clusters is to utilize the Bayesian Information Criterion (BIC) [22]. The BIC is defined in Equation (16).

$$BIC = -2L(\Theta_M) + K_M ln(N) \tag{16}$$

Where $L(\Theta_M)$ is the total log-likelihood function computed at the optimum and K_M is the number of independent parameters of a GMM of size M. The EM algorithm is implemented in a loop for a GMM where $M = 1:10$ and the BIC value is calculated for each mixture model. The mixture model with the lowest BIC is chosen as it will have the highest likelihood and least complexity. $M = 1:10$ was chosen as it was assumed unlikely that there would be more than 10 unique traffic routes in the region surrounding S_0.

Multiple Trajectory Prediction

Given the unique clusters of trajectories formed from the GMM, the future position of the target vessel can be estimated. The trajectory clusters are in this study considered as discrete possible future paths a selected vessel may follow. Therefore, a trajectory prediction with respect to each cluster is conducted, resulting in a prediction of M possible trajectories the selected vessel may follow. The algorithm utilized to achieve each individual prediction is presented in this section.

The prediction algorithm utilized in this study is the same as that utilized in [6]. This method is based on the work in [13] where a Single Point Neighbor Search Method for vessel trajectory prediction based on historical AIS data was presented.

The prediction algorithm is an iterative process where the estimated future state of the observed vessel for iteration k is defined as $\hat{S}_k$:

$$\hat{S}_k \rightarrow [\hat{\phi}_k, \hat{\lambda}_k, C\hat{O}G_k, S\hat{O}G_k, \hat{T}_k] \tag{17}$$

The predicted position in state $\hat{S}_k$, is determined using the position data in the previous state $\hat{S}_{k-1}$. The position $[\hat{\phi}_k, \hat{\lambda}_k]$ is estimated as a distance ΔL from $[\hat{\phi}_{k-1}, \hat{\lambda}_{k-1}]$ in the direction of $C\hat{O}G_{k-1}$. This is visualized in Figure 2. $\hat{T}_k$ is calculated in (18).

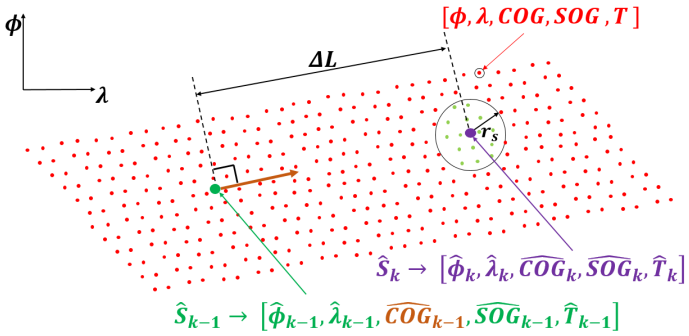

FIGURE 2. ILLUSTRATION OF PREDICTION ALGORITHM.

$$\hat{T}_k = \hat{T}_{k-1} + \frac{\Delta L}{\hat{SOG}_{k-1}} \tag{18}$$

$\hat{COG}_k$ and $\hat{SOG}_k$ are determined using a clustering technique in the original $\phi - \lambda$ vector space. The distances to all points are calculated and all points within a distance corresponding to a predefined radius, r_s are defined as the cluster C_k for $\hat{S}_k$. This is visualized in in Figure 2. $\hat{COG}_k$ and $\hat{SOG}_k$ are calculated in Equations (19) and (20) as the median values of the cluster.

$$\hat{COG}_k = median(C_k[COG]) \tag{19}$$

$$\hat{SOG}_k = median(C_k[SOG]) \tag{20}$$

The prediction algorithm then repeats until $\hat{T}_k \geq T_{pred}$. This results in a predicted trajectory for the selected vessel from T_0 and the subsequent period corresponding to T_{pred}. The predicted trajectory is then linearly interpolated at 30 second intervals from T_0 up to and including T_{pred}.

General Algorithm

The complete algorithm described in the previous sections is summarized in this section and consists of the steps below. It produces M predicted trajectories for the selected vessel.

1: Initialize S_0
2: Initialize initial cluster
3: Extract trajectories from initial cluster
4: Generate trajectory feature vectors using PCA
5: **for** M = 1:10 **do**
6: Cluster trajectories using GMM
7: Calculate BIC
8: **end for**
9: Choose model M s.t. BIC is minimized
10: Classify trajectories to M clusters
11: **for** i = 1:M **do**
12: Run trajectory prediction algorithm on trajectory cluster i
13: **end for**

RESULTS AND DISCUSSION

As outlined in the Methodology section of this paper, features are generated to represent each trajectory extracted from the initial cluster centered around S_0. Two randomly selected vessel states in the AIS data, denoted Vessel A and Vessel B, were chosen to illustrate the results of the algorithm in this section.

Data Set

The historical AIS data set utilized to achieve the results in this section was provided by the Norwegian Coastal Administration. One year of AIS data, from the 1^{st} of January, 2017 to the 1^{st} of January, 2018 for the region around the city of Tromsoe, Norway was utilized in this study. This data set corresponds to approximately 15 million AIS data points.

GMM Trajectory Clustering

Principle Components. Figure 3 illustrates the projection of the trajectory vector data onto the subspace spanned by the three largest eigenvectors (e_1, e_2, e_3), or principle components, when running the algorithm for a random initial state S_0 in the AIS data set. Each row and column represent the i^{th} principle component. Along the diagonals are the kernel density estimates using Gaussian kernels for each class in the GMM. Each color represents a unique class. It should be noted that these density functions are not normalized by the prior distributions. The remaining subplots represent the pairwise scatter plots of the projected data.

Using only the three dimensions comprising the largest three principle components, the GMM has been able to successfully cluster the trajectories into three distinct clusters for the case of Figure 3. As previously discussed, these vectors represent the highest degree of variation between the trajectory feature vectors. It was found for this particular case that the three largest eigenvalues explained 98% of the variance in the data. As such, the choice of three eigenvectors can be a good representation of the initial AIS data set that was considered for this analysis. Furthermore, these eigenvectors will preserve 98% of the total variance of the same data set. Given that the original trajectory feature vector contained position, course and speed values for every data point along the trajectory, the principle components will describe the largest difference in position, speed and course along their

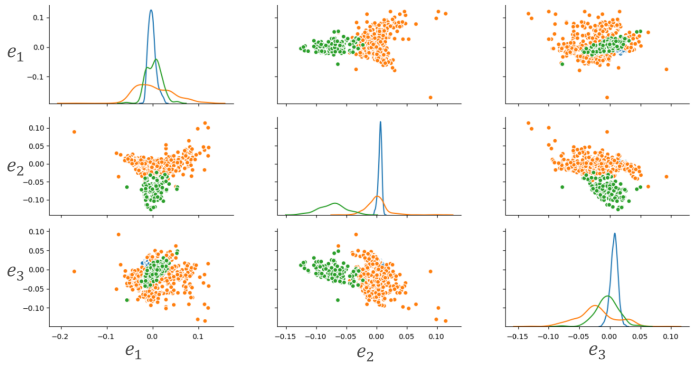

FIGURE 3. PROJECTED TRAJECTORY VECTOR DATA ONTO SUBSPACE SPANNED BY PRINCIPAL COMPONENTS.

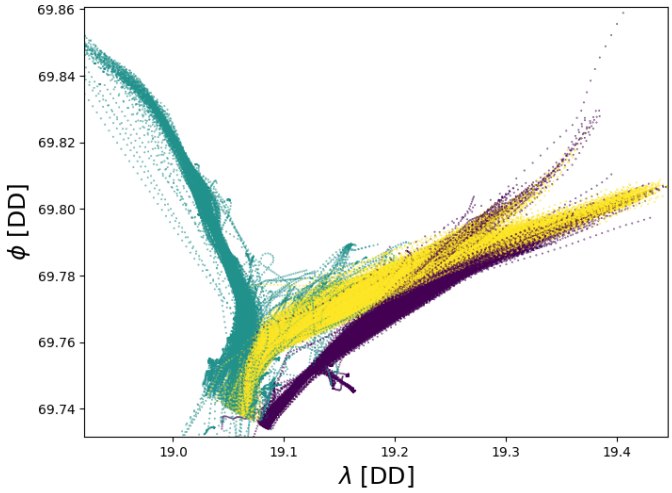

FIGURE 4. RESULT OF UNSUPERVISED GAUSSIAN MIXTURE MODELING OF VESSEL TRAJECTORIES FOR VESSEL A.

axes by projecting all the high dimensional trajectory data onto the three principle components.

Vessel A. Figure 4 illustrates the results of the GMM clustering for Vessel A. In this case, three distinct clusters due to traffic separation are observed. It appears that the clustering algorithm performed quite well based on visual inspection of the spacial data presented in the figure. The yellow and purple clusters have similar trajectories, but the algorithm has nonetheless been able to distinguish them using the principle components of the trajectory data.

Vessel B. Figure 5 illustrates the results of the GMM clustering for Vessel B. In this case, the clustering has resulted

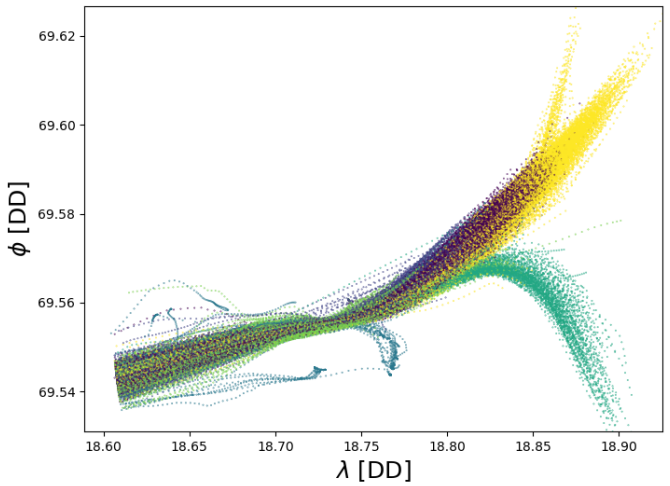

FIGURE 5. RESULT OF UNSUPERVISED GAUSSIAN MIXTURE MODELING OF VESSEL TRAJECTORIES FOR VESSEL B.

in a higher number of underlying models in the GMM, where a total of six classes were discovered. Compared to Vessel A, the extracted trajectories for Vessel B have a much higher degree of similarity for the majority of the of the trajectory in terms of positional data. There are two main traffic lanes illustrated by the green cluster and yellow cluster. The remaining four clusters, however, have similar trajectories to that of the yellow.

In this case, it is likely the principle components have accounted for differences in the speed of the vessels. This results in the four remaining clusters ending much earlier. This property of the PCA based GMM shows that the algorithm also can discover trajectories with similar spatial properties, but that sail the same traffic route with various speeds. This effect is of course not solely based on the projection of the *SOG* data, but also the positional data at various time instances. The *COG* data will in this particular case not account for much of the variation between all classes aside from the green class culminating in the lower right corner of the figure.

Multiple Trajectory Prediction

Vessel A. Figure 6 illustrates the predicted trajectories for each cluster discovered by GMM. In this case, the algorithm was only able to predict two trajectories, despite there in fact being three clusters present. This is due to the fact that the initial data points for the purple cluster in Figure 4 were too far away from S_0 in the $\phi - \lambda$ plane to create a cluster in the initial phase of the prediction algorithm. Physically, this is logical, as the vessel would have to make an extreme course change in order to enter the traffic route represented by the purple cluster, and as such a prediction along this route is unlikely.

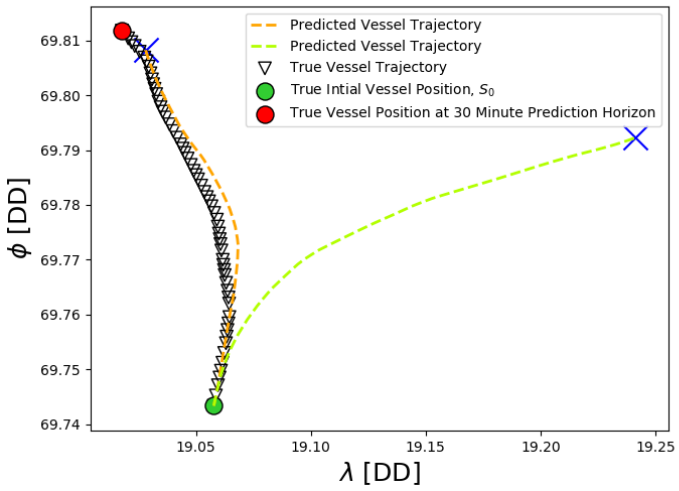

FIGURE 6. MULTIPLE TRAJECTORY PREDICTION RESULTS FOR VESSEL A.

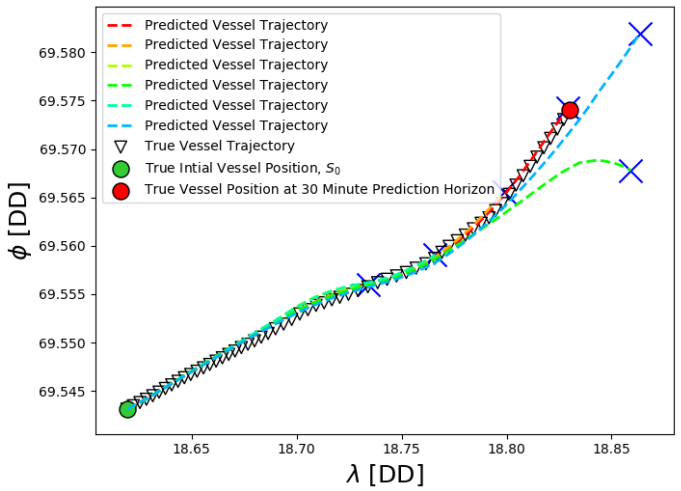

FIGURE 7. MULTIPLE TRAJECTORY PREDICTION RESULTS FOR VESSEL B.

The two predicted trajectories correspond to predictions run on the green and yellow cluster data in Figure 4. It is clear that Vessel A in fact follows the route corresponding to the green cluster. The prediction along this route follows the actual trajectory of the Vessel A reasonably well, and has a predicted position at a 30 minute prediction horizon quite close to the actual position of the Vessel A after 30 minutes.

Vessel B. In the case of Vessel B, trajectory predictions for all six clusters present are realized when running the algorithm. The six predicted trajectories are illustrated in Figure 7. The predicted trajectory corresponding to the red line in Figure 7 appears to almost perfectly predict the trajectory of Vessel B. The 30 minute prediction coincides nearly exactly with the true position of Vessel B after 30 minutes. This is likely due to the discovery of a cluster of vessel trajectories with a similar speed profile to that of Vessel B.

Prediction Accuracy

In order to evaluate the accuracy of the algorithm, it was run 100 times on randomly chosen data points in the AIS data set. Each randomly chosen point was defined as S_0 for that run of the algorithm. Given that the algorithm generates multiple predicted trajectories, only one will correspond to the cluster that best fits the true vessel trajectory. As such, the predicted trajectory with the lowest average estimated position error was chosen to evaluate the performance of the algorithm. Figure 8 illustrates the median error of the 100 runs at various prediction horizons for these trajectories.

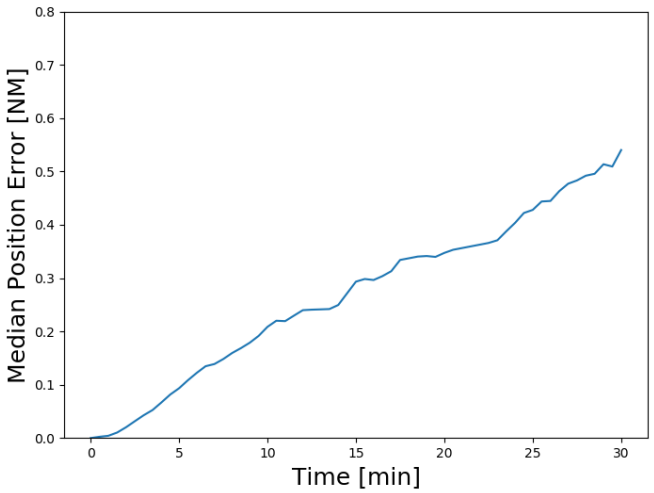

FIGURE 8. MEDIAN POSITION ERROR OF 100 RUNS.

Upon investigation it was found that a few runs resulted in much higher position errors than the majority. This was likely due to low data density in the areas corresponding to the randomly chosen S_0. In this case, the algorithm is unable to effectively predict the trajectory of the vessel. Therefore, the results of these runs can be considered outliers. As such, it was deemed more appropriate to present the median error than the root mean square error as it is sensitive to outliers. Nonetheless, this is a weakness of the algorithm, as it only performs well in areas with high data density.

Generally, it appears that the algorithm performs quite well. The trend observed in Figure 8 is also to be expected, as the extracted trajectory data diverges as time increases from S_0. A shortcoming of the algorithm is that it has no way of indicating which cluster is the most likely. As such, each prediction is considered equally likely. If a future vessel were to utilize such an algorithm for collision avoidance purposes, it would need to evaluate the collision risk with respect to all predicted trajectories.

CONCLUSION AND FURTHER WORK

The algorithm presented in this study results in multiple predicted trajectories based on the state of a vessel for a specified prediction horizon. By extracting trajectories from an initial cluster centered about the selected vessel state, the algorithm is able to identify multiple possible routes the vessel may follow. Using Principle Component Analysis, the algorithm generates a lower dimensional feature vector for each trajectory. These feature vectors are then input to unsupervised Gaussian mixture modeling resulting in the clustering of the trajectories. Subsequently, a prediction algorithm is applied to the data in each cluster resulting in a prediction for each possible route.

The results indicate that the algorithm is quite effective in clustering the trajectories, where at least one cluster corresponds closely to the actual trajectory of the vessel. The predicted trajectories of the best cluster have a high degree of accuracy for prediction horizons up to 30 minutes, but the algorithm suffers in areas of low AIS data density. Certain runs of the algorithm also result in a failed prediction for some clusters due to the proximity of the initial data points of the cluster with respect to the vessel state. The algorithm also has no way of identifying the most likely cluster with respect to the vessel state, resulting in all predictions being considered equally likely. Nonetheless, the algorithm outlined in this study should allow future vessels to predict the future trajectory of a target vessel quite well, as one of the clusters identified will provide an accurate prediction.

Any geographic region is potentially a candidate for this approach, depending on the density of the historical AIS data. In open waters, the algorithm will likely function quite well, as there is little variation in traffic routes. In more congested areas, such as coastal regions and port zones, the traffic picture naturally becomes more complex with a higher number of possible routes the vessel may follow, and the resultant predictions more uncertain.

The approach also suffers as it does not account for variation in weather conditions, as all AIS trajectories are considered in the prediction irrespective of the prevailing weather conditions. Further work will investigate combining historical metocean data with the historical AIS data to enhance the predictions, as the prevailing environmental conditions will affect the ship maneuvers. Further work will also include a method to classify the vessel state to one of the discovered clusters. This approach will provide a probability associated with the state belonging to each cluster, and as such a probability of the resultant trajectory prediction. Methods to quantify the probability of a position for a given prediction horizon along a predicted trajectory will also be investigated. Alternative trajectory prediction methods run on each cluster will also be evaluated with respect to the accuracy of the predictions.

ACKNOWLEDGMENT

This work was supported by the Norwegian Ministry of Education and Research and the MARKOM2020 project, a development project for maritime competence established by the Norwegian Ministry of Education and Research in cooperation with the Norwegian Ministry of Trade and Industry. The authors would also like to express gratitude to the Norwegian Coastal Administration for providing access to their AIS database.

REFERENCES

[1] Endsley, M. R., 1995. "Toward a Theory of Situation Awareness in Dynamic Systems". *Human Factors: The Journal of the Human Factors and Ergonomics Society,* ***37***(1), pp. 32–64.

[2] Perera, L. P., Ferrari, V., Santos, F. P., Hinostroza, M. A., and Guedes Soares, C., 2015. "Experimental Evaluations on Ship Autonomous Navigation and Collision Avoidance by Intelligent Guidance". *IEEE Journal of Oceanic Engineering,* ***40***(2), apr, pp. 374–387.

[3] Tam, C., Bucknall, R., and Greig, A., 2009. "Review of collision avoidance and path planning methods for ships in close range encounters". *Journal of Navigation,* ***62***(3), pp. 455–476.

[4] Perera, L. P., Oliveira, P., and Guedes Soares, C., 2012. "Maritime Traffic Monitoring Based on Vessel Detection, Tracking, State Estimation, and Trajectory Prediction". *IEEE Transactions on Intelligent Transportation Systems,* ***13***(3), sep, pp. 1188–1200.

[5] Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., and Huang, G.-B., 2017. "Exploiting AIS Data for Intelligent Maritime Navigation: A Comprehensive Survey From Data to Methodology". *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–24.

[6] Murray, B., and Perera, L. P., 2018. "A Data-Driven Approach to Vessel Trajectory Prediction for Safe Autonomous Ship Operations". In Proceedings of The Thirteenth International Conference on Digital Information Management (ICDIM), pp. 240–247.

[7] Gunnar Aarsæther, K., and Moan, T., 2009. "Estimating Navigation Patterns from AIS". *Journal of Navigation,* ***62***(04), oct, pp. 587–607.

[8] Dobrkovic, A., Iacob, M.-E., and Van Hillegersberg, J., 2018. "Maritime pattern extraction and route reconstruction from incomplete AIS data". *International Journal of Data Science and Analytics,* ***5***(2), Mar, pp. 111–136.

[9] Pallotta, G., Vespe, M., and Bryan, K., 2013. "Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction". *Entropy,* ***15***(12), jun, pp. 2218–2245.

[10] Xiao, F., Ligteringen, H., Van Gulijk, C., and Ale, B., 2015. "Comparison study on AIS data of ship traffic behavior". *Ocean Engineering,* ***95***, pp. 84–93.

[11] Zissis, D., Xidias, E. K., and Lekkas, D., 2016. "Real-time vessel behavior prediction". *Evolving Systems,* ***7***(1), mar, pp. 29–40.

[12] Mazzarella, F., Arguedas, V. F., and Vespe, M., 2015. "Knowledge-based vessel position prediction using historical AIS data". In 2015 Sensor Data Fusion: Trends, Solutions, Applications (SDF), IEEE, pp. 1–6.

[13] Hexeberg, S., Flaten, A. L., Eriksen, B.-O. H., and Brekke, E. F., 2017. "AIS-based vessel trajectory prediction". In 2017 20th International Conference on Information Fusion (FUSION), IEEE.

[14] Dalsnes, B. R., Hexeberg, S., Flåten, A. L., Eriksen, B.-O. H., and Brekke, E. F., 2018. "The neighbor course distribution method with gaussian mixture models for ais-based vessel trajectory prediction". In 2018 21st International Conference on Information Fusion (FUSION), IEEE, pp. 580–587.

[15] Li, H., Liu, J., Liu, R., Xiong, N., Wu, K., and Kim, T.-h., 2017. "A Dimensionality Reduction-Based Multi-Step Clustering Method for Robust Vessel Trajectory Analysis". *Sensors,* ***17***(8), aug, p. 1792.

[16] Laxhammar, R., Falkman, G., and Sviestins, E., 2009. "Anomaly detection in sea traffic-a comparison of the gaussian mixture model and the kernel density estimator". In Information Fusion, 2009. FUSION'09. 12th International Conference on, IEEE, pp. 756–763.

[17] Karhunen, K., 1946. "Zur spektraltheorie stochastischer prozesse". *Annales Academiae Scientiarum Fennicae,* ***37***.

[18] Jollife, I., 1986. "Principle Component Analysis". *Springer-Verlag*.

[19] Diamantaras, K. I., and Kung, S. Y., 1996. *Principal component neural networks: theory and applications*, Vol. 5. Wiley New York.

[20] Hyvärinen, A., 2009. Principal component analysis.

[21] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., 2000. "Speaker verification using adapted gaussian mixture models". *Digital signal processing,* ***10***(1-3), pp. 19–41.

[22] Schwarz, G., et al., 1978. "Estimating the dimension of a model". *The annals of statistics,* ***6***(2), pp. 461–464.