Unsupervised Flight Phase Recognition with Flight Data Clustering based on GMM

Datong Liu[1,2], Ning Xiao[2], Yujie Zhang[1,2], Xiyuan Peng[1,2]
[1]School of Electronics and information Engineering, Harbin Institute of Technology, Harbin 150080, China
[2]Department of Automatic Test and Control, Harbin Institute of Technology, Harbin 150080, China
Email: {liudatong, ningxiao, hnhyzyjlh, pxy}@hit.edu.cn

***Abstract*—Currently, with the rapid development of the aviation industry, researchers are paying more attention to the improvement of aviation safety. Aviation safety mainly includes flight safety, aviation ground safety, and air defense safety. In terms of flight safety, the analysis of large amounts of flight data has gradually become a useful tool for timely detection of potential dangers at various stages of flight. As a result, flight data analysis has been one of the hot topics in aviation. However, due to the complexity of the aircraft operating conditions throughout the aircraft, if the data is analyzed at the entire flight phase, it is very difficult and time consuming to identify the problematic fight phase. Therefore, flight phase recognition for civil aircraft is implemented in this study. A flight phase recognition method based on Gaussian Mixture Model (GMM) is proposed in this work, which is the important foundation for timely detecting the abnormal event and improving the system safety and reliability. Firstly, the FDR data are preprocessed by spline interpolation and normalization, and then a GMM-based flight phase clustering is realized. In addition, a set of evaluation method is developed to evaluate the quality of flight phase recognition result. Finally, the effectiveness of the method is verified by using real FDR data from NASA's open database.**

***Keywords*—flight data, flight phase recognition, Gaussian mixture model, flight safety, aircraft**

I. Introduction

In recent years, with the rapid development of local economies and airlines, the proportion of civil aircraft in domestic travel tools has gradually increased. As a convenient transportation, aircraft has been accepted and selected by more and more people. Moreover, aviation safety has become the most important issue for various countries and aviation organizations [1], [2], including aviation ground safety, air defense safety and flight safety. Among them, in terms of flight safety, one of best tools for detecting the potential dangers in time at various phases of flight is analyse for large amounts of flight data. Therefore, the analysis of aircraft flight data has become one of research hotspots in aviation.

Currently, most aircraft use Flight Data Recorder (FDR) to record flight information of aircraft during flight. FDR data refer to data obtained by encoding and decoding these flight information recorded by FDR [3]. They include the health-related parameters of various components of the aircraft, such as engine oil temperature, rotor speed, pitch angle, flight altitude and so on. As a result, FDR data can be analyzed the of flight accidents improve the airline safety [1]. Moreover, in order to further improve safety, it is necessary to adopt a more active approach. In this case, abnormal detection and risk identification with fight data have become the main research topics in aviation area.

For example, Li et al. [4] developed a new data-driven method that uses a clustering algorithm to detect anomalous flight, which was suitable for a variety of standard operations. In order to detect potential security problems in large databases of commercial aircraft, Das et al. [5] proposed an anomaly detection algorithm based on kernel learning theory, which treated flights as a cluster, compresses important features of each flight and then identifies flight level anomalies. Matthews et al. [6] used the multiple kernel anomaly detection algorithm to identify potentially dangerous major events in the National Airspace System. Song et al. [7] used a least squares support vector machine to implement multi-step prediction for univariate time series, and adopted a prediction strategy for anomaly detection. Das et al. [8] proposed a feature extraction algorithm for symbolic dynamic filtering, which can be directly applied to perform fault detection and classification. It can found more flight anomalies, improved the performance of the algorithm, and had the characteristics of flexibility. Pang et al. [9] proposed a real-time anomaly detection method based on the combination of the improved anomaly detection and mitigation and Gaussian process regression (GPR) , which realized the incremental detection of future data samples. The model is more effective than Multilayer Perception. Li et al. [10] proposed a flight data analysis method based on Gaussian mixture model clustering. It evaluated flight risk by whether it had a common mode, and performed abnormal data mode detection without pre-specifying the search content and can accurately detect flight anomalies.

In summary, many methods have been proposed for aircraft anomaly detection to improve the safety of aircraft. However, most of them directly use clustering, extraction of feature values and other algorithms to detect outliers. If the impact of flight mode switching is not fully considered, which will take a lot of time to determine whether the outliers have operational significance. At present, the flight phase of the aircraft is mainly identified by the following methods: regression-based classifiers, Supervised learning and filtering methods based on

This study was partially supported by National Natural Science Foundation of China under Grant No. 61803121, 61571160 and 61701131, and ROOT-CLOUD Experiment Test and Validation Environment of Industrial Internet Platform Supported by 2018 Innovation and Development Project of Industrial Internet.

978-1-7281-4460-3/20/$31.00 ©2020 IEEE

variable constraints of each stage and substage. For instance, Sun et al. [11] used a combination method of density-based spatial clustering of Applications with noise algorithm and fuzzy logic recognition to recognized the flight phase of the aircraft, which can effectively process a large number of scattered flight data. Chin et al. [12] carried out experiments on various flight phase recognition methods and considered that the sliding window regression method has high recognition accuracy. Yang et al. [13] applied fuzzy support vector machine to aircraft flight motion recognition, and its recognition rate was significantly improved compared with traditional support vector machine.

Up to now, supervised learning is a commonly used learning method in flight phase recognition research. However, much recorded flight data are unlabeled and cannot be analyzed using supervised learning methods. Therefore, the focus of this study is to recognize the flight phase using the unsupervised learning algorithm referring to Gaussian mixture model (GMM). GMM is a classical unsupervised learning algorithm whose essence is the parameter probability density function, expressed as the weighted sum of gaussian distribution density [14]. In theory, any data sample can be represented by a sufficient linear combination of Gaussian distributions. Result of GMM-based flight phase recognition can be used in flight data records of two types of aircraft, labelled data and unlabelled data, providing a basis for subsequent aircraft anomaly detection.

The rest of this study is organized as follows. Section II introduces the related principle of GMM. Section III describes the basic framework of flight phase recognition method based on GMM. Section IV introduces the data used in the experiment, some experimental details and experimental results. In addition, the accuracy of flight phases recognition is evaluated in section IV. Section V summarizes this study and discusses future work.

II. Gaussian Mixture Model

A. Definition of Gaussian Distribution

Gaussian distribution is a very common continuous probability distribution in probability theory. It is often used to represent an unknown random variable [15]. When variate $\mathbf{x}$ represents unidimensional data, the probability density function subject to the Gaussian distribution can be expressed by,

$$p(\mathbf{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mathbf{x}-\mu)^2}{2\sigma^2}}, \tag{1}$$

where μ is the mean or expectation of the distribution, which determines the location of the distribution. And σ is the standard deviation, which can change the magnitude of the distribution.

When variate $\mathbf{x}$ represent multivariate, the probability density function subject to the Gaussian distribution is expressed as,

$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{(\mathbf{x}-\mu)^T \sum^{-1} (\mathbf{x}-\mu)}{2}}, \tag{2}$$

where D is the dimension of the sample vector. μ is the mean or expectation of the distribution and Σ is the covariance. Moreover, $|\Sigma|$ represents the determinant of Σ and Σ^{-1} means the inverse matrix of Σ.

B. Gaussian Mixture Models

The GMM is a parameter probability density function, which is an extension of a single Gaussian probability density function. It can be expressed as a weighted sum of Gaussian distribution densities [16].

The advantage of GMM is that it is a soft clustering method, in which the sample points get the probability of belonging to each class rather than a definite classification mark. The GMM with the K components can be expressed by,

$$p(\mathbf{x}|\lambda) = \sum\nolimits_{i=1}^{K} \omega_i g(\mathbf{x}|\mu_i, \Sigma_i), \tag{3}$$

where $\mathbf{x}$ are the M-dimensional vectors. $\lambda_i = \{\omega_i, \mu_i, \Sigma_i\}$ means the parameters of the GMM. ω_i, $i = 1, ..., K$ are the weight of the Gaussian mixture model which should satisfy $\omega_i \geq 0$ and $\sum_{i=1}^{K} \omega_i$=1. μ_i, $i = 1, ..., K$ are the mean of vectors and Σ_i, $i = 1, ..., K$ represent the covariance matrixes of Gaussian. The mixed Gaussian component density can be expressed by,

$$p(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)' \sum_i{}^{-1} (\mathbf{x}-\mu_i)}. \tag{4}$$

In order to build a GMM model, the expectation-maximization (EM) algorithm is usually used to estimate the parameters of the GMM [17]. The parameters of the GMM should be updated in each interaction to obtain likelihood convergence [18]–[20]. Using EM algorithms to estimate GMM parameters typically includes three steps.

Firstly, initializing the means μ_i , covariances Σ_i and mixing weights ω_i , and evaluating the initial value of the log likelihood.

Secondly, also known as E step, calculating the posterior probability $\gamma_j(\mathbf{x})$ based on the current parameter value [21]. It can be written as :

$$\gamma_j(\mathbf{x}) = \frac{\omega_k g(\mathbf{x}|\mu_k, \Sigma_k)}{\sum\limits_{j=1}^{K} \omega_j g(\mathbf{x}|\mu_j, \Sigma_j)}. \tag{5}$$

Thirdly, also known as M step, re-estimating the parameters using the current responsibilities. The parameters of GMM are updated using the following equations.

$$\mu_i = \frac{\sum\limits_{n=1}^{N} \gamma_j(x_n) x_n}{\sum\limits_{n=1}^{N} \gamma_j(x_n)}, \tag{6}$$

$$\Sigma_j = \frac{\sum\limits_{n=1}^{N} \gamma_j(x_n) (x_n - \mu_j) (x_n - \mu_j)^T}{\sum\limits_{n=1}^{N} \gamma_j(x_n)}, \tag{7}$$

$$\omega_j = \frac{1}{N}\sum_{n=1}^{N}\gamma_j(x_n), \tag{8}$$

where N is the number of samples points.

In the EM algorithm for GMM, the log likelihood is:

$$\ln p(\mathbf{x}|\mu,\Sigma,\pi) = \sum_{n=1}^{K} ln\{\sum_{k=1}^{K}\omega_k g(x_n|\mu_k,\Sigma_k)\}. \tag{9}$$

Repeat step 2 and 3 until (9) converges, then parameters for each Gaussian component can be obtained.

III. Framework

GMM-based flight phase recognition is mainly divided into four steps, as shown in Fig. 1.

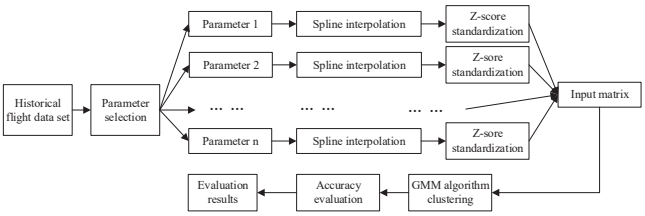

Fig. 1. The framework of GMM-based method for flight phase recognition.

Firstly, the flight data are characterized by the spline interpolation, and the sampling data of different frequency is processed by the same length. Secondly, the flight parameters after the equal length processing are standardized by Z-score standardization to eliminate the impact of raw data at different orders of magnitude. Then, the aircraft flight phases are recognized using the GMM algorithm, and the clustering results are evaluated using real FDR data. The specific process is shown as follows.

A. Parameter selection

Through the analysis of experimental data, a total of 7 parameters were selected as input to the GMM model in this study, including baro correct altitude, pitch angle, roll angle, true heading angle, altitude change rate, power lever angle, and engine fan speed. Their specific information is shown in Table **??**.

TABLE I
INPUT PARAMETER TABLE

Number	The full name	sampling frequency	Unit
1	Baro correct altitude	4	ft
2	Pitch angle	8	deg
3	Roll angle	8	deg
4	True heading angle	4	deg
5	Altitude change rate	4	ft/min
6	Power lever angle	4	deg
7	Engine fan speed	4	%rpm

B. Cubic spline interpolation

Cubic spline interpolation is used to solve the problem of different parameter lengths in this study. Furthermore, compared with other interpolation methods, the spline interpolation method has the advantages of good convergence and stable calculation process [22].

Cubic spline interpolation given a set of n+1 data points (x_i, y_i) , and any two values of x_i is not the same. Moreover,$a = x_0 < x_1 < \cdots < x_n = b$, the spline interpolation function $S(x)$ satisfies the following condition:

- $S(x) \in C^2[a,b]$
- $S(x)$ is a polynomial of degree 3 on each subinterval $[x_{i-1}, x_i]$,where $i = 1, \cdots, n$.
- $S(x_i) = y_i$, for all i=0, $1 \cdots, n$.

$S(x)$ assumed satisfies the (10).

$$S(x) = \begin{cases} C_1(x), & x_0 \le x \le x_1 \\ \cdots \\ C_i(x), & x_{i-1} < x \le x_i \\ \cdots \\ C_n(x), & x_{n-1} < x \le x_n \end{cases} \tag{10}$$

where each $C_i = a_i + b_i x + c_i x^2 + d_i x^3 (d_i \neq 0), i = 1, \cdots, n$ is a cubic function.

If $S(x)$ want to be determined, a_i , b_i , c_i and d_i for each i are determined by:

- $C_i(x_i - 1) = y_{i-1}$ and $C_i(x_i) = y_i$, $i = 1, \cdots, n$.
- $C_i{}'(x_i) = C_{i+1}{}'(x_i), i = 1, \ldots, n-1$.
- $C_i{}'(x_i) = C_{i+1}{}''(x_i), i = 1, \ldots, n-1$.

$4n$ conditions need to be determined to solve the polynomial, however, it just have $4n-2$ conditions [23]. So boundary conditions should be added to solve this problem. The three boundary conditions are shown as follows:

Firstly, boundary conditions called clamped boundary conditions and the first derivative of the boundary point is known:

$$C_1{}'(x_0) = f_0{}', C_n{}'(x_n) = f_n{}' \tag{11}$$

Secondly, it is second derivatives at the endpoints are known:

$$C_1{}''(x_0) = f_0{}'', C_n{}''(x_n) = f_n{}'' \tag{12}$$

When $C_1''(x_0) = C_n''(x_n) = 0$, it is called simple boundary conditions.

Thirdly, if function $f(x)$ is a periodic function with $x_n - x_0$, $S(x)$ is also a periodic function with period $x_n - x_0$. So

$$C_1(x_0) = C_n(x_n), C_1{}'(X_0) = C_n{}'(x_n) \tag{13}$$

$$C_1''(x_0) = C_n''(x_n) \tag{14}$$

If $S(x)$ meet third boundary condition, spline interpolation are called periodic splines.

C. Standardized processing

Because the range of parameter value in calculating distance has significant influence on the relative weight of feature, the Z-score standardized method is used for FDR data before using GMM to recognize flight phase of an aircraft.

Data standardization is an important process for converting data into same order of magnitude. Z-score is a commonly used method of data standardization. The basic Z-score formula for a simple can be expressed by,

$$x' = \frac{x - \mu}{\sigma}, \tag{15}$$

where μ is the mean of the simple and σ is the standard deviation of the sample.

D. Recognizing flight mode with GMM

According to the different characteristics of flight data in different flight phases, GMM is used to cluster the pre-processed flight data, which can output the flight phases corresponding to each flight data segment more accurately. In the following experiments, the aircraft flight phases are divided into four phases: taxiing, climbing, cruising, and approaching.

IV. Experimental Results and Analysis

A. FDR data description

The FDR datasets used in this study are from the open database of NASA's official website [24]. The sample provides actual flight data for 35 different types of regional aircraft at different starting points and destinations over three years. While the files contain detailed aircraft dynamics, system performance, and other engineering parameters, all information traceable to a particular airline or manufacturer has been erased. There are 186 parameters monitored in the flight sample, including parameter types: continuous, discrete and alarm signals.

A total of 55 flight data are used as experimental samples and one of the samples are used as a case for specific analysis in this study. Due to the FDR data is that the sampling frequency of the sensor is different, resulting in unequal data lengths of the parameters. Moreover, the FDR data used in this study is because a physical meaning of each parameter is different, each parameter has a quite different value range [24]. In addition, the FDR data used in this study come from different routes, so there is no direct comparability between flights.

The original data curves of the seven parameters are shown in Fig. 2.

From Fig. 2, it can be seen that the original flight data have different lengths and different orders of magnitude. Therefore, the data are preprocessed by spline interpolation and z-score standardization before they are used in flight phases recognition. The result of pre-processing is shown in Fig. 3.

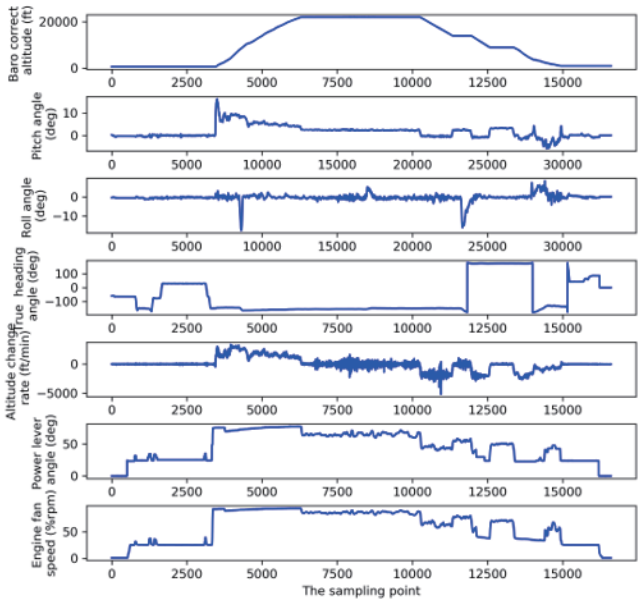

Fig. 2. Original curves of the selected flight parameters.

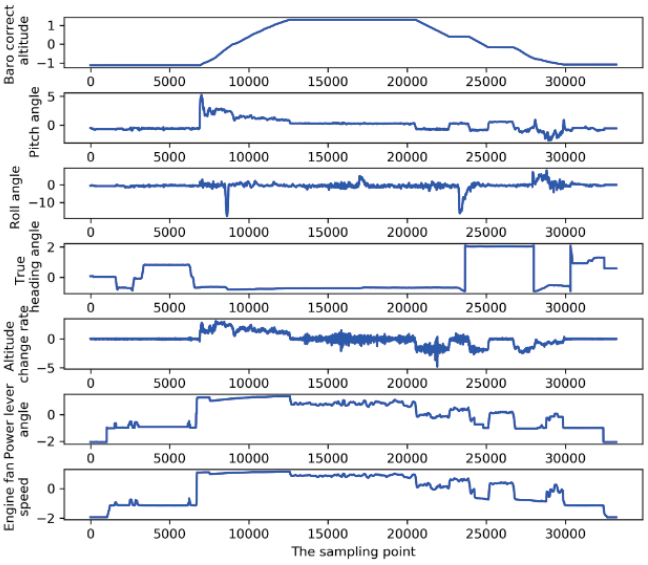

Fig. 3. Data preprocessing results of flight data.

B. Recognizing flight phase with GMM

In this study, four flight phases of aircraft taxiing, climbing, cruising and approaching are considered. The flight data of Flight 686 on April 13, 2001 at 14:37 is used as an example. The results of flight phase recognition by GMM are shown in Figs. 4 and 5.

From Fig. 4, it can be seen that the flight phase of each sampling point. Furthermore, in order to express the recognition results of flight phases more intuitively, the results of flight phases recognition are reflected in the barometric altitude figure and it can be shown in Fig. 5.

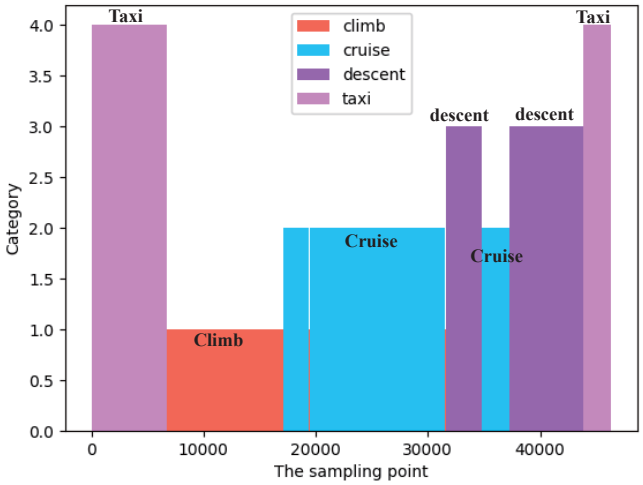

Fig. 4. Statistical results of flight phases recognition.

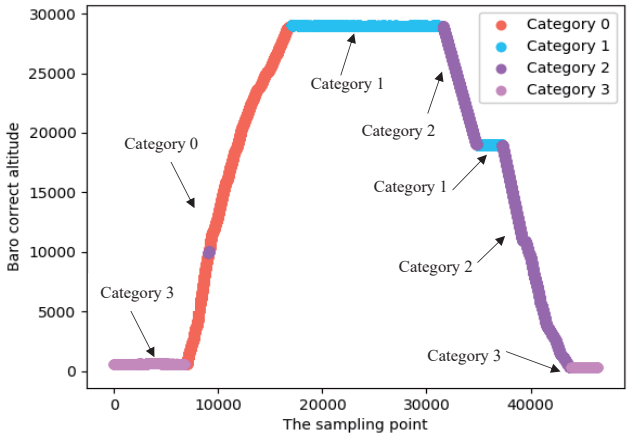

Fig. 5. the recognition results on barometric altitude figure.

C. Evaluation Criteria for Recognition of flight phases

In most cases, the clustering algorithm does not know the real labels of data samples, which is an unsupervised learning algorithm. Therefore, it is impossible to evaluate the clustering algorithm completely based on the evaluation criteria of the supervised learning method. However, flight data samples used in the experiment has a status word called PH at different stages of the reaction aircraft. Therefore, this experiment uses the PH status word as the real label of the data to analyze the clustering experiment results.

The flight phases of the aircraft are divided into four stages, hence, the PH value was converted into four corresponding phases, and the result is shown in the Fig.6.

In Fig. 6, number 1 to 4 means taxi, climb, cruise and approach. In this study, the data of 55 flights are analyzed, and the status word called PH is used as the real label to calculate the recognition accuracy of flight phases. The results are shown in the Table II.

In this experiment, the highest accuracy of flight phase

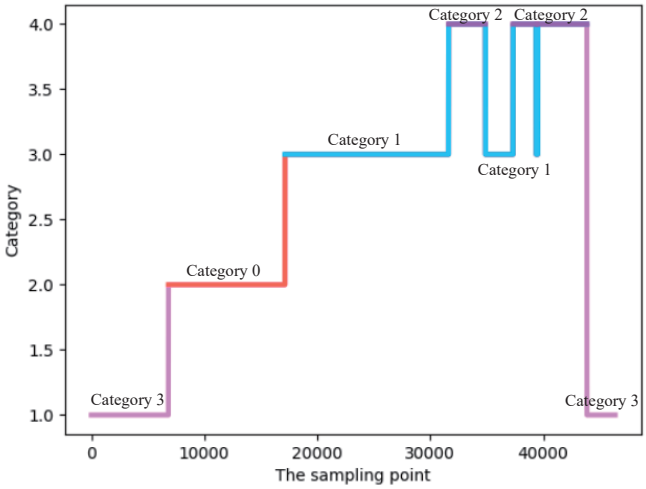

Fig. 6. A parametric data graph called PH.

TABLE II
STATISTICS OF FLIGHT PHASES RECOGNITION RESULTS

Accuracy	60%∼70%	70%∼80%	80%∼90%	90%∼100%
Quantity of flight sampes	2	4	16	33

recognition is 98.52%, the lowest is 62.77%, and the average accuracy is about 90.04%. A statistical figure of this results is shown as follow Fig. 7.

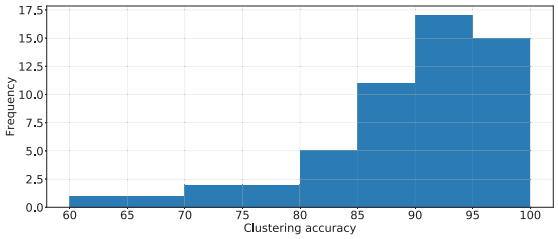

Fig. 7. Statistical results of flight phases identification of 55 experiments.

V. CONCLUSIONS

In order to fully consider the influence of flight mode switching in aircraft anomaly detection more, this study proposes a aircraft flight phase identification method based on GMM. To verify the performance of the flight phase recognition method based on GMM, A set of experiments are performed. the real flight data come from official NASA open database. In addition, a set of evaluation criteria are used to evaluate the accuracy of flight phase recognition. For the 55 flight samples tested, the average accuracy of state recognition is up to 90.04%. The experimental results show that the GMM algorithm has a good performance on recognizing flight phase of aircraft.

Although the GMM can be used to recognize the flight phase of the aircraft, its performance and stability is easily affected by the difference in flight data. In our future work, we will focus on optimizing the GMM to further improve the stability of flight phase recognition.

REFERENCES

[1] L. Li, "Anomaly detection in airline routine operations using flight data recorder data," Ph.D. dissertation, Massachusetts Institute of Technology, 2013.

[2] Y. Zhang, L. Wang, S. Wang, P. Wang, H. Liao, and Y. Peng, "Auxiliary power unit failure prediction using quantified generalized renewal process," *Microelectronics Reliability*, vol. 84, pp. 215–225, 2018.

[3] T. M. Mcdade, "Advances in flight data acquisition and management systems," in *Digital Avionics Systems Conference, 1998. Proceedings., 17th DASC. The AIAA/IEEE/SAE*, 1998.

[4] L. Li, S. Das, R. John Hansman, R. Palacios, and A. N. Srivastava, "Analysis of flight data using clustering techniques for detecting abnormal operations," *Journal of Aerospace information systems*, vol. 12, no. 9, pp. 587–598, 2015.

[5] S. Das, B. L. Matthews, and R. Lawrence, "Fleet level anomaly detection of aviation safety data," in *2011 IEEE Conference on Prognostics and Health Management*. IEEE, 2011, pp. 1–10.

[6] B. Matthews, A. N. Srivsatava, J. Schade, D. R. Schleicher, K. Chan, R. Gutterud, and M. Kiniry, "Discovery of abnormal flight patterns in flight track data," in *2013 Aviation Technology, Integration, and Operations Conference*, 2013, p. 4386.

[7] S. Ge, L. Jun, D. Liu, and Y. Peng, "Anomaly detection of condition monitoring with predicted uncertainty for aerospace applications," in *2015 12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, vol. 1. IEEE, 2015, pp. 248–253.

[8] S. Das, S. Sarkar, A. Ray, A. Srivastava, and D. L. Simon, "Anomaly detection in flight recorder data: A dynamic data-driven approach," in *2013 American Control Conference*. IEEE, 2013, pp. 2668–2673.

[9] J. Pang, D. Liu, H. Liao, Y. Peng, and X. Peng, "Anomaly detection based on data stream monitoring and prediction with improved gaussian process regression algorithm," in *2014 International Conference on Prognostics and Health Management*. IEEE, 2014, pp. 1–7.

[10] W. Zhao, F. He, L. Li, and G. Xiao, "An adaptive online learning model for flight data cluster analysis," in *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018, pp. 1–7.

[11] J. Sun, J. Ellerbroek, and J. Hoekstra, "Flight extraction and phase identification for large automatic dependent surveillance–broadcast datasets," *Journal of Aerospace Information Systems*, pp. 566–572, 2017.

[12] H.-J. Chin, A. Payan, C. Johnson, and D. N. Mavris, "Phases of flight identification for rotorcraft operations," in *AIAA Scitech 2019 Forum*, 2019, p. 0139.

[13] J. Yang and S.-S. Xie, "Fuzzy support vector machines based recognition for aeroplane flight action," *Acta Aeronautica et Astronautica Sinica*, vol. 26, no. 6, pp. 738–742, 2005.

[14] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[15] N. D. Socci, D. D. Lee, and H. S. Seung, "The rectified gaussian distribution," in *Advances in neural information processing systems*, 1998, pp. 350–356.

[16] D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 659–663, 2008.

[17] D. H. H. Santosh, P. Venkatesh, P. Poornesh, L. N. Rao, and N. A. Kumar, "Tracking multiple moving objects using gaussian mixture model," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 3, no. 2, pp. 114–119, 2013.

[18] L. Li, R. J. Hansman, R. Palacios, and R. Welsch, "Anomaly detection via a gaussian mixture model for flight operation and safety monitoring," *Transportation Research Part C: Emerging Technologies*, vol. 64, pp. 45–57, 2016.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[20] Z.-K. Huang and K.-W. Chau, "A new image thresholding method based on gaussian mixture model," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 899–907, 2008.

[21] N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall, "Computing gaussian mixture models with em using equivalence constraints," in *Advances in neural information processing systems*, 2004, pp. 465–472.

[22] S. McKinley and M. Levine, "Cubic spline interpolation," *College of the Redwoods*, vol. 45, no. 1, pp. 1049–1060, 1998.

[23] C. Habermann and F. Kindermann, "Multidimensional spline interpolation: Theory and applications," *Computational Economics*, vol. 30, no. 2, pp. 153–169, 2007.

[24] C.-H. Lee, H.-S. Shin, A. Tsourdos, and Z. Skaf, "Data analytics development of fdr (flight data recorder) data for airline maintenance operations," in *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2017, pp. 289–294.