Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/117/2025.19947

A Survey of Pilot Behavior Detection Methods

Yue Lin[1,a,*]

[1]*Dalian University of Technology, Liaoning, 116024, China*
a. yl782@student.le.ac.uk
**corresponding author*

Abstract: This research investigates the use of action and gesture detection technology to analyze pilot behavior and ensure flight safety by employing artificial intelligence. In civil aviation, pilot and passenger safety is of paramount importance ,and the pilot behavior is an important factor in safety. Thus, it is necessary to detect the pilot behavior to secure the flight process which involves the detection of multiple modalities. In the past, most researchers have focused on single-modal behavioral detection methods, primarily gestures. However, a single modality, which may also include physiological data like EEG, is not a good and comprehensive understanding of a series of scenarios. Therefore, this paper proposes using multimodal data fusion to integrate behavior detection data from multiple sources. This paper summarizes and outlines the main methods of pilot behavior detection, such as cockpit voice, gesture detection and text information, and look forward to the possibility of multimodal data fusion of these there modalities data.

Keywords: Cockpit voice, gesture detection, behavior detection, deep learning, text recognition.

1. Introduction

With the development of China's economy and aviation technology, civil aircraft has become the preferred mode of travel. As the industry evolves, in order to reduce the cost of pilot configuration, researchers from various countries to explore the pilot configuration of the research began to shift from DPO to SPO or even unmanned. Under this trend, the safety of flight has also been emphasized. According to previous studies and statistics, human factors account for about 60% to 80% of flight accidents, and about 70% of the flight accidents in the past 10 years were caused by flight crew factors [1]. Therefore, detecting the behavioral state and operational actions of pilots is a feasible measure to avoid safety accidents, and the behavioral detection of pilots under SPO needs to be carried out by multiple sensors forsingle data recognized by a single sensor does not allow for a comprehensive detection and determine the pilot's behavioral state. This paper focuses on the behavioral detection methods of pilots based on the acquisition of three different modal data.

2. Visual gesture detection

For the study of human hand movements, gesture detection, as a key step in gesture recognition, is not ultimately aimed at detecting gestures in isolation, though it is a basis for other applications. Gesture detection obtains the start and end time of each gesture from a sequence of consecutive gestures to track human behaviour. The popular gesture detection methods are skin color information,

DOI: 10.54254/2755-2721/117/2025.19947

depth information, geometric information and motion information [3][4]. Skin color is an important feature, gesture detection based on skin color features is usually performed by building a color space model followed by skin color detection. For instance, models based on skin color, such as the YCrCb model by Chai and Habibi [5], help segment hand movements in complex backgrounds. Geometric information-based approaches use contour tracking, with algorithms like the Canny operator and Snake model [7][8], though these are computationally demanding. In 1987, Kass, Witkin and Terzopoulos [7] proposed the parametric model Snake [8], which is often called the parametric active contour model, to locate the edge of the image by parametric curves from the energy point of view. Taking the Snake model as an example, the model's sensitivity to noise and contrast is utilized to achieve target contour tracking under complex backgrounds. The algorithm has its limitations: when the initial contour of the Snake model has a large gap with the target contour, the algorithm is computationally intensive. With the development of depth sensors, there are also gesture detection methods based on depth information, such as using RGB-D cameras for depth segmentation or combining color information for segmentation. The difficulty of hand gesture detection is hand segmentation, compared with traditional image processing hand segmentation, deep learning based hand segmentation methods have the advantage of automatic learning, for example, Kankana Roy et al. combined hand skin color features and neural network algorithms to effectively improve the accuracy and stability of hand segmentation and recognition tasks [9].

Vision-based gesture behavior analysis systems usually involve several processes of gesture detection, gesture classification, motion tracking, and behavior understanding and description. Since the pilot's operation tasks rely heavily on hand movements and the cockpit environment is complex, it is necessary to analyze the depth information in order to judge the pilot's hand operation.

The goal of the gesture detection task is interpreted as obtaining the position of the gesture in the time domain, based on the definition of Detection in the work of Molchanov et al. on gesture detection and recognition [15]. The goal of gesture detection is to obtain the start and end time of each gesture from a sequence of consecutive gestures. Earlier detection methods partially detected the start and end of gestures with manual feature thresholds, e.g., Peng et al. detected gestures in terms of the distance offset between the hand position in each gesture frame and the hand position in the static template [16], and Chai et al. detected gestures in terms of the maximum height of both hands [10].

The first step of the detection method based on depth information is to acquire depth information through depth camera, and then gesture detection after 3D reconstruction of the cockpit. And there are two main methods to acquire depth information, through binocular stereo vision and through depth camera.

Although binocular stereo vision technology has the advantage of using only natural light to acquire information and is highly adaptable. However, it requires a large amount of computation, and in real flights, there may be situations where only instrumented lights are available and there is a lack of natural light, in which case binocular vision technology cannot be effective.

The depth camera can avoid the above problems and will be more suitable to get the depth information.

2.1. Selection of depth camera

This section will introduce the three depth cameras Leap motion, Kinect, and RealSense and select them based on their respective suitability in the cockpit.

The Leap Motion controller is a small gesture acquisition device from LEAP currently used in virtual reality and sign language interpretation. Leap Motion can accurately track each finger of the operator's hand, but Leap Motion's effective detection range is limited to 25 to 600 mm,making it unsuitable for cockpit environments (Figure 1).

DOI: 10.54254/2755-2721/117/2025.19947

Figure 1: Leap motion controller disassembly diagram

The Kinect sensor (Figure2)is a body sensing device from Microsoft that works well with the Kinect for Window SDK in acquiring color images, depth images, and skeletal data of the human body. However, it is too bulky to fit properly in a cockpit environment.

Figure 2: Kinect Schematic

RealSense(figure 3) is a small form factor depth camera from Intel.RealSenseR200 is small and has a field of view of 0.7-3.3m, both of which meet the requirements of the cockpit. The relevant parameters can refer to Table 1

Figure 3: Realsense R200

Table 1: Realsense series depth camera parameter indicators and configuration environment list

	Realsense R200		Realsense SR300	
	color video streaming	deep video streaming	color video streaming	deep video streaming
Maximum effective accuracy	1940 x 1080	640 x 480	1940 x 1080	640 x 480
refresh rate	30fps	30fps	30fps	30fps
volumetric	130mm*20mm*7mm		110mm*12.6mm*4.1mm	
Working distance	0.7~4m	0.7~4m	0.2m~1.5m	0.2m~1.5m
Required Operating Systems	Windows 8.1 or 10 64-bit		Windows 10 64-bit	

DOI: 10.54254/2755-2721/117/2025.19947

2.2. Hand behavior detection network

The research task of target detection is to obtain the accurate location information and category information of specified objects in an image or video. Famous network models include R-CNN, SSD, YOLO series network and so on. This part will focus on R-CNN network, YOLO series network.

In 2014, RossGirshick et al designed a region-based target detection network to construct a two-stage target detection implementation[13]. The R-CNN network structure is shown in Figure.4 For an input image, the R-CNN target detection network uses a selective search algorithm to generate about 2,000 rectangular candidate box subregions with different sizes domains in the input image, with different rectangular candidate box subregions selecting different components of the target.

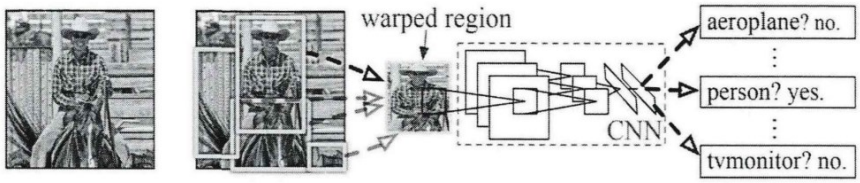

Figure 4: Overall structure of R-CNN network [12]

Compared to R-CNN, the YOLO family of algorithms offers faster speeds and allows end-to-end training and prediction, making it user-friendly. However, the detection accuracy is low compared to that for R-CNN.

According to the data organized by Zhiren Xiao based on the dataset, YOLOv5 among the YOLO series of algorithms has some advantages in both detection speed and accuracy [14]. The network structure of YOLOv5 is shown in Figure 5 below.

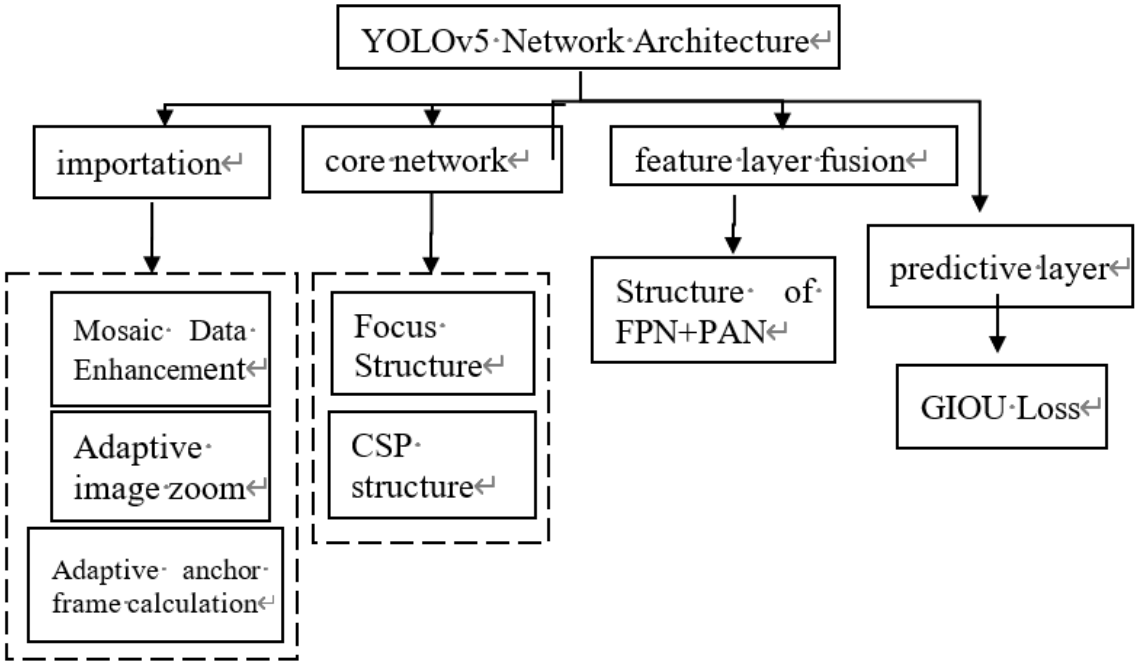

Figure 5: YOLO v5 Network Architecture Diagram

Input Layer: Preprocesses image data to enhance the model's performance.

Backbone: Uses the Focus and CSP modules to boost computational efficiency and learning capacity while reducing the computational load.

DOI: 10.54254/2755-2721/117/2025.19947

Feature Fusion Layer (Neck): Combines FPN (Feature Pyramid Network) and PAN (Path Aggregation Network) structures to integrate feature maps of varying scales, improving prediction accuracy for images of different resolutions.

Prediction Layer: The final layer converts network output into prediction frames and class probabilities. Using GIOU Loss improves the model's accuracy and robustness in target detection.

3. Cockpit voice

Recent research has developed detection methods for noisy environments. For instance, Li et al. proposed a short-time energy speech detection algorithm that operates effectively under low SNR conditions, making it suitable for diverse noisy settings. Zhu Xinyang et al. introduced a robust speech endpoint detection method specifically for cockpits, enabling accurate detection of pilot speech in such a challenging environment [10][11].

Pilot's voice information is also an important basis for analyzing pilot's behavior or actions, and the pilot's voice can be extracted from the voice recordings using Voice Activity Detection (VAD) technique, which finds out the beginning and end of the speech from a specified speech signal. The subsequent Feature Extraction is performed by removing redundant information such as noise and extracting the recognizable components of a speech. Using some extracted feature vectors for the keywords that need to be detected, a template is trained and the keywords are detected using the template.

3.1. Text structuring

For cockpit speech, the pilot's behavior can be analyzed by converting the speech signal into quantifiable text. Therefore, it is necessary to use a speech recognition model to structure the speech signal into text.

DeepSpeech2, an end-to-end speech recognition model based on deep learning proposed by Baidu with 2016, can directly convert speech signals into text strings, simplifying a series of complex processes of model training. However, as cockpit communication often involves technical terms, generic models may struggle with accuracy. To address this, Lei Zhongzhou developed a cockpit-specific speech dataset, and the text of the recording was selected from the standard Boeing 737-800 pilot standard communication phrases, a total of 20 entries(Table 2). The trained model converts cockpit speech into text accurately.

Table 2: Pilot Standard communication phrases

NO.	Standard communication phrases	NO.	Standard communication phrases
1	Pre-flight checklist	11	Pressurization mode selector to AUTO
2	Pre-start checklist	12	Window heating ON
3	Pre-taxi checklist	13	Autobrake armed
4	Pre-takeoff checklist	14	Landing gear UP
5	Post-takeoff checklist	15	Inertial navigation system selector OFF
6	Descent checklist	16	Engine bleed air ON
7	Approach checklist	17	Anti-collision lights ON
8	Landing checklist	18	Air conditioning pack to AUTO
9	Shutdown checklist	19	Probe heating ON
10	Disembarkation checklist	20	Isolation valve to AUTO

DOI: 10.54254/2755-2721/117/2025.19947

After confirming the above text, the dataset production is carried out and the process is shown in Figure 6.

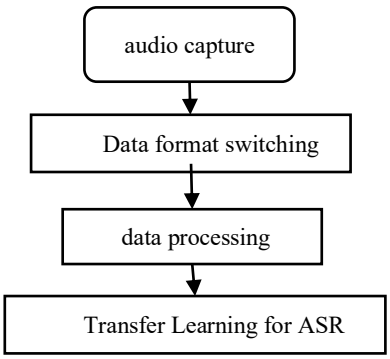

Figure 6: Flowchart of data production

The trained model is then used to convert the speech signal into the corresponding text string.

3.2. Keyword detection

In land and air communications , the use of speech keyword detection technology can quickly locate the keywords in the speech of the call and obtain the desired information, which is shown as figure 7. Deep neural networks (DNNs) are particularly suited for cockpit keyword detection due to their real-time performance.

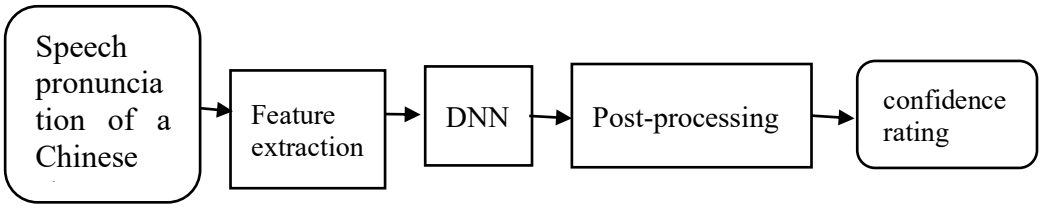

Figure 7: Flowchart of Keyword detection

Huang Xianghong proposed residual gating neural network as a deep neural network for cockpit speech keyword detection (figure 8)technology for this purpose[16]. The civil aviation call recordings are firstly feature extracted, and then the residual gating neural network is utilized for keyword detection.

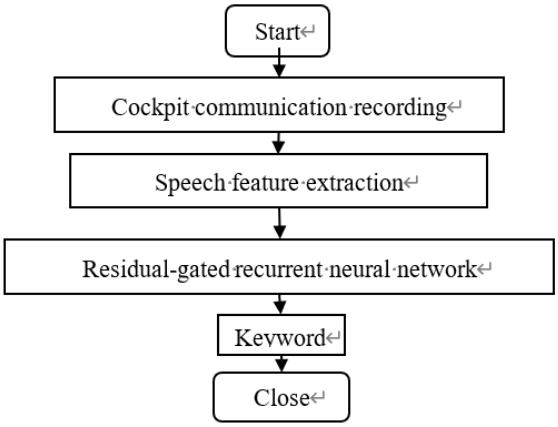

Figure 8: Flow chart of civil aviation speech keyword detection

DOI: 10.54254/2755-2721/117/2025.19947

The current model with the best effect of speech keyword detection in general-purpose scenarios is the residual network model, however, the residual network can't capture the contextual information well, Huang proposed residual gated recurrent neural network based on the residual network [16]. The structure of residual gated recurrent neural network is given as figure 9.

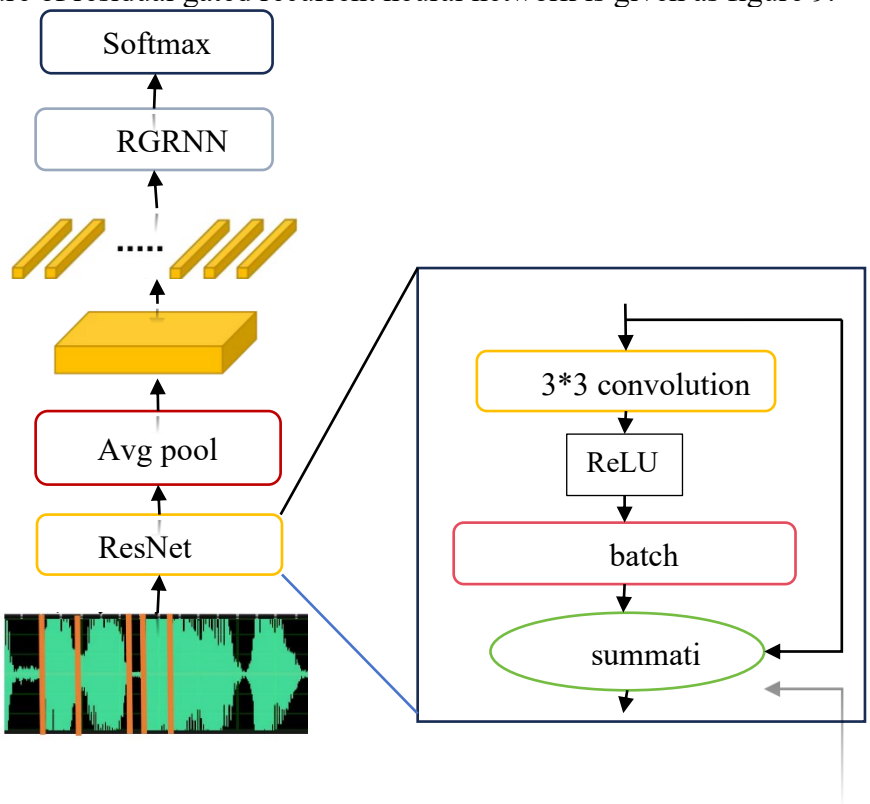

Figure 9: Structure of the residual network model [16]

At the residual network layer, the convolutional layer extracts the feature sequence from each frame of audio features. Above the residual network layer is an average pooling layer whose main function is to average the feature maps extracted from the residual network layer, which can act as a down sampling layer, and then above the average pooling layer is a gated recurrent network layer, which can be used to make predictions for each frame of the feature sequence. Finally, the Soft max layer is used to take the prediction of each frame from the gated recurrent network layer as an input and output the keyword category to which the whole audio belongs.

4. Text message recognition

Text detection, mainly OCR (Optical character recognition) technology, the current commonly used Chinese OCR recognition engine ABBYY and Tesseract [12]. ABBYY can scan paper documents, images, PDF and other documents can be edited documents, ABBYY converts paper documents, images, and PDFs into editable text, while Tesseract offers flexibility for custom training and recognition of specific fonts. This adaptability makes Tesseract effective for recognizing checklist text in flight management computers (FMCS).

Pilots perform a series of checks on pre-flight operations using checklists during pre-takeoff preparations and enter textual information into the Flight Management Computer (FMCS), and therefore obtaining such textual information can be another method of detecting pilot behavior.

For the text information in the checklist(figure 10), which usually contains text in the image data, it is necessary to first locate the text content in the image and then recognize it. There are two methods

DOI: 10.54254/2755-2721/117/2025.19947

to localize the text, one is based on the edge detection method [18] by using the difference between the region to which the text belongs and the background of the image, identify the edge of the text region, which has the advantage of fast recognition speed, but in the face of the complexity of the background of the image, the accuracy is poor, the other is based on the detection of artificial intelligence technology [19], with the help of the current target detection algorithms such as improved RCNN, YOLO algorithm, etc., to detect the text region as a target. And the image background in the checklist is not complicated, so the text content localization can be achieved by using the edge detection method.

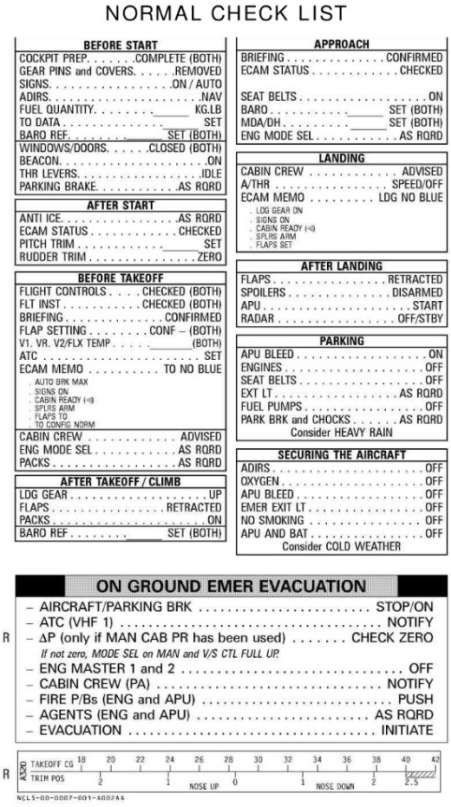

Figure 10: Airbus A320 Normal Checklist (English)

The text recognition task is divided into printed text recognition and handwritten text recognition. The checklist used by pilots belongs to the former, and the recognition method used by the former is called optical character recognition (OCR), which was proposed by the German scholar Tausheck in 1929 and has been widely used [20], while the recognition method adopted by the latter utilizes related artificial intelligence methods such as convolutional neural networks [21].To recognize text from image data in checklists, edge detection and AI-based methods can be employed. Edge detection, which identifies text region edges, works well for simpler backgrounds, such as those in cockpit checklists. OCR, first proposed by German scholar Tausheck in 1929 and later developed by Hewlett-Packard, is effective for printed text recognition, making it suitable for flight checklist data.

Tesseract for Hewlett-Packard experiments in 1985 developed out of the OCR recognition engine, to 1995 to become one of the OCR industry's most accurate text recognition engine, can be used as a flight checklist for text recognition tools.

DOI: 10.54254/2755-2721/117/2025.19947

5. Future outlook

Multimodal fusion technology is the integration of data information from multiple sensors for a comprehensive understanding of the scene. It has applications in many aspects, for example, for human gesture detection in computer vision technology, Cheng Zhang proposed a graph-convolutional gesture recognition method based on multimodal skeleton[16], while Zhongzhou Lei researched for the technical processing of multimodal decision-making system for artificial intelligence co-pilot[17].

Currently, no multimodal fusion technique exists for comprehensive pilot behavior assessment by combining voice, gesture, and text information. Future studies can analyze the data of the three modalities and select appropriate multimodal data fusion methods [22], such as decision-level fusion, encoder fusion, or multimodal information alignment methods to achieve multimodal data fusion, so as to overcome the shortcomings of the single data, offering a more comprehensive view of pilot behavior.

6. Conclusion

This article reviews three different modalities for pilot behavior detection, introduces the basic principles and network models of each behavior detection method, and analyzes the environment for behavior detection in the cockpit. While voice, gesture, and text are vital for behavior detection, integrating physiological data like EEG can further improve accuracy.In addition, multimodal fusion is an important issue in pilot detection and has great potential. Continuous innovation in multimodal detection can support the shift from DPO to SPO configurations, ensuring safety and operational efficiency in future aviation technology.

References

[1] *WANG Lei, LIANG Yan. Statistics and analysis of global civil aviation accident investigation data[J]. China Transportation review, 2021, 43(3): 7-12.*

[2] *YANG Kun, WANG Haoran. Behavior pattern recognition of pilots using head up dispaly[J]. Science Technology and Engineering, 2018, 18(29): 226- 231.*

[3] *Fujimura K, Nanda H. Visual tracking using depth data: IEEE Computer Society, US 7372977 B2 [P]. 2008.*

[4] *Francke H, Ruiz-Del-Solar J, Verschae R. Real-Time Hand Gesture Detection and Recognition Using Boosted Classifiers and Active Learning[M]. Advances in Image and Video Technology. Springer Berlin Heidelberg, 2007: 533-547.*

[5] *Chai D..Ngan K. N. . Locating facial region of a head-and-shoulders color image. in:Proceedings of the 3rd International Conference on Automatic Face and Gesture Rocognition, Nara Japan, 1998, 124-129.*

[6] *Bastos R, Sales Dias M. Skin color profile capture for scale and rotation invariant hand gesture recognition[C]// International Gesture Workshop. Springer, Berlin, Heidelberg, 2007:81-92.*

[7] *Zass M, Witkin, D. Terzopoulos D. Snakes: active consour models. ijcv, 1(4):321-331, 1988.3.*

[8] *M kass, A. Witkin, D Terzo Polous. Snake:Active Contour Models [C].Proceeding of the 1st International Conference on computer Vision. London:IEEE Computer Society Press.1987, 257-268*

[9] *Roy K, Mohanty A, Sahay R R. Deep learning based hand detection in cluttered environment using skin segmentation[C]//Proceedings of the IEEE international conference on computer vision workshops, 2017:640-649.*

[10] *Q. Li, J. S. Zheng, A. Tsai, et al. Robust endpoint detection and energy normalization for real-time speech and speaker recognition[J]. IEEE Transactions on Speech and Audio Processing, 2002, 10(3): 146-156*

[11] *ZHU Xinyang, HUANG Dan, LU Yanyu, et al. Pilot Speech Endpoint Detection in Aircraft Cockpit Noisy Environment[J]. Computer Engineering, 2018, 44(01):317-321.*

[12] *Smith R. An overview of the Tesseract OCR engine[C]. International Conference on Document Analysis and Recognition(ICDAR), 2007, 2:629-633*

[13] *Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate for accurate object. Detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.*

[14] *Xiao Zhiren, Zhang Chenhao , Wang Shoyu. A gesture detection system based on YOLOv5s network model[J] Information Record Material, 2023, 24(02):183-185+188.DOI:10.16009/j.cnki.cn13-1295/tq.2023.02.004*

[15] *Lei Zhongzhou. Research on Multimodal Decision-making Technology for Audio-Visual Perception in Artificial Intelligence Co-polit for flight[D]. Civil Aviation Flight University of China, 2023.DOI:10.27722/d.cnki.gzgmh. 2023.000031.*

[16] *Huang Xianghong. Study on Keyword Spotting Method for Radiotelephony Communication in Civil Aviation Based on Deep Learning[D]. Nanjing University of Aeronautics and Astronautics, 2022.DOI:10.27239/d.cnki.gnhhu.2022. 001101.*

[17] *Zhang Cheng. Research on computer vision based detection and recognition of dynamic human gestures[D]. Beijing Institute of Technology, 2023.DOI:10.26935/d.cnki.gbjgu.2023.000367.*

[18] *Sun Hua, Zhang Hang. A review of Chinese character recognition methods [J]. Computer Engineering, 2010, 36 (20): 194-197.*

[19] *Tang Minli, Xie Shaomin, Liu Xiangrong. Detection and Recognition of Handwritten Texts in Ancient Water Book Based on Faster-RCNN [J]. Journal of Xiamen University (Natural Science Edition), 2022, 61(2): 272-277.*

[20] *Wang Xuguang, Yin Shige. Research and Practice of OCR Technology in Enterprise Document Recognition [J]. Information and Computer (Theoretical Edition), 2022, 34(18): 175-178.*

[21] *Zhao Fan, Zhang Lin, Wen Zhiquan, et al. A Direct and Efficient Approximation and Localization Method for Chinese Characters in Natural Scenes [J]. Computer Engineering and Applications, 2021, 57(6): 159-167.*

[22] *ZHANG Hucheng, LI Leixiao, LIU Dongjiang. A survey of multimodal data fusion research [J/OL].Journal of Frontiers of Computer Science and Technology https://link.cnki.net/urlid/11.5602.tp.20240620.1752.002*