Dean&Francis

ISSN 2959-6157

Pilot Operation Action and Attitude Recognition Technology Based on Multi-View Visual Image Sequence

Zexi Qiao[1,*]

[1]School of Physics and Mechatronic Engineering, Hubei University of Education, Wuhan, 430205, China
*Corresponding author: q19308652235@outlook.com

Abstract:

Pilot maneuver and posture recognition serves as a key technical cornerstone for supporting quantitative evaluation of flight training and cockpit safety early warning. Its recognition accuracy directly determines the objectivity of training effect scoring, while its response speed is linked to the timeliness of cockpit risk early warning, making it of great significance for improving the quality of aviation training and flight safety standards.Multi-view vision technology, leveraging the unique advantages of non-contact measurement and accurate 3D scene reconstruction, has gradually replaced traditional solutions such as single-view observation and sensor-based contact monitoring, becoming the core technical support in this field. It consists of three core modules: image preprocessing, key point detection, and temporal modeling. Image preprocessing can eliminate light and clutter interference in the cockpit; key point detection is able to locate critical movement parts like hands, arms, and torso; and temporal modeling connects continuous operations to form an action logic chain.Relying on these modules, it can real-time correct non-standard operations such as push rod force deviation and lever position offset during flight training, and quickly identify dangerous postures like excessive forward leaning and delayed hand response in safety monitoring. Meanwhile, this technology needs to address severe occlusion caused by control levers and instrument panels in the cockpit, as well as the strict real-time requirement of millisecond-level response, while lightweight algorithm optimization and multi-modal data fusion provide a clear path for its subsequent development.

Keywords: Pilot maneuver recognition; Multi-view vision; temporal modeling

ISSN 2959-6157

1. Introduction

In flight training, traditional evaluation methods relying on instructors' subjective judgments struggle to accurately quantify subtle operational variations like 1° joystick deflection, resulting in unscientific assessment of training effectiveness. In cockpit scenarios, dangerous postures such as pilots staring at instruments for extended periods may lead to safety incidents if not promptly warned. Pilot operation recognition technology has become crucial for enhancing flight training quality and ensuring cockpit safety. Multi-view vision technology, with its non-contact measurement capabilities, precise 3D coordinate reconstruction, and occlusion resistance, meets the operational demands of pilot control environments, providing vital support for industry development.

Early single-lens vision systems, unable to resolve three-dimensional scale ambiguity, failed to meet the precision requirements for pilot posture and motion recognition, leading to their gradual obsolescence.Subsequently, multi-camera vision technology emerged, achieving three-dimensional keypoint localization through camera calibration and stereo matching, thereby overcoming the limitations of monocular vision. In recent years, the integration of deep learning with image sequence analysis has become the mainstream approach, focusing on temporal modeling of dynamic movements to enable complex operational process recognition. Current research priorities include optimizing multi-camera configurations and calibrations for compact cockpits, developing keypoint tracking algorithms for occlusion scenarios, extracting spatio-temporal fusion features, and creating lightweight models that ensure real-time performance.

This paper will systematically elaborate on the technical principles of multi-view vision and the requirements for pilot motion recognition. It will outline core methodologies including preprocessing of multi-view visual image sequences, keypoint detection and tracking, feature extraction, and recognition algorithms through a technical workflow. Case studies will be analyzed from human-machine interaction scenarios such as flight training evaluation and cockpit safety monitoring. Existing challenges in technology-scenario adaptation will be summarized, while future research directions are projected.

2. Technical Basis

2.1 Core Principle of Multi-Eye Vision Technology

2.1 1. Calibrate-Correct-Stereo Match

The multi-view visual system achieves 3D reconstruction through a "calibration-correction-solid matching" workflow, as shown in Fig. 1 [1]. In cockpit scenarios where space is limited, the Zhang Zhengyou calibration method requires optimization and can be adapted using folded calibration plates to accommodate spatial constraints. The SGBM (Partial Global Block Matching) algorithm, known for its strong resistance to light interference, effectively addresses cockpit instrument glare issues and is widely applied in such environments. By calculating disparity maps and combining them with camera parameters to convert them into 3D coordinates, this method provides a geometric foundation for recognizing pilots' movements and postures [2].

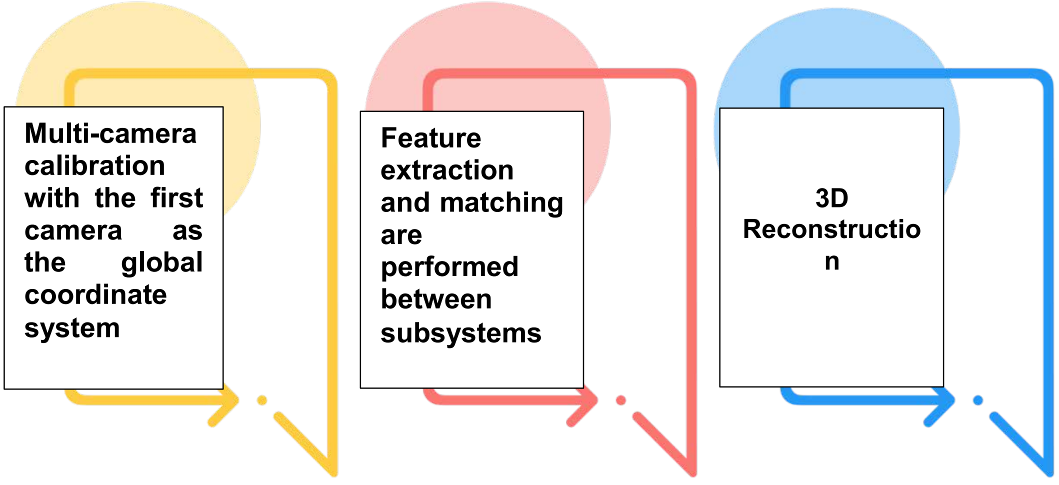

Fig. 1 Calibration-Correction-Stereo Matching Process

2.1.2 Comparison with other technologies

Unlike inertial sensors, multi-camera vision eliminates the need for pilot-mounted equipment, thus avoiding operational interference. Compared to LiDAR systems, multi-camera vision cameras cost only 1/10 to 1/3 of the price while being more adaptable for engineering deployment. The three key advantages of multi-camera vision—non-contact operation, cost-effectiveness, and 3D precision-make it the preferred technology for scenarios where pilots cannot wear distracting devices and cabin budgets are limited.

2.2 Key Indicators of Pilot Action/Pose Recognition

To accurately distinguish between subtle maneuvers and posture changes, the hand and head pitch angle deviation control stick deflection must maintain within a specific range. Any deviations may lead to misjudgment of operational compliance or fatigue status, thereby compromising flight training evaluations and safety monitoring accuracy. Flight training assessments require real-time feedback with latency under 100ms, whereas cockpit safety monitoring must issue instant alerts for hazardous postures within 200ms to enable timely interventions and prevent accidents. The system must adapt to complex cabin environments—including instrument glare (light intensity fluctuations ±30%), arm obstruction rates under 30%, and variations in pilot attire—to ensure uninterrupted detection processes and reliable identification results [3,4].

3.Core Methods

3.1 Multi-View Visual Image Sequence Preprocessing

The cockpit features a compact space (typical width <1.5m), employing a "1 front + 2 side" camera layout to cover upper body and hand control areas. Due to spatial constraints, a "step-by-step calibration + joint optimization" approach is adopted: first calibrating each camera individually, then optimizing external parameters through co-view feature points to reduce calibration error to.To address frame rate differences among cameras, spatio-temporal synchronization of image sequences is achieved through interpolation. Motion blur caused by rapid pilot operations is mitigated using adaptive exposure control [1,4].

3.2 Pilot Human Key Point Detection and Tracking

Liu Hao and his team proposed an image brightness adjustment and order-switching attention module to construct a lightweight keypoint detection network integrating HRNet and Transformer [5], capable of extracting 59 limb and hand keypoints. Extensive experiments have validated the effectiveness of this method, with the overall architecture for pilot limb keypoint detection shown in Figure 2 [6]. The Kalman filter's "predict-update" mechanism is widely used to enhance temporal continuity in keypoint trajectories due to its noise suppression capability. Its core concept originates from established applications in uniform motion target tracking—— such as the "Kalman Filter-Based Target Recognition Tracking and Shooting System Design," which constructs a uniform motion model to describe coordinate and velocity relationships through state equations, achieving trajectory optimization via observation fusion [7]. This framework can be adapted for pilot action tracking: In scenarios like steady joystick rotation and stable head posture changes, keypoint coordinates and velocities serve as state vectors, while multi-camera visual detection results act as observations. Dynamic adjustment of noise covariance suppresses detection jitter caused by cockpit reflections and occlusions.

ISSN 2959-6157

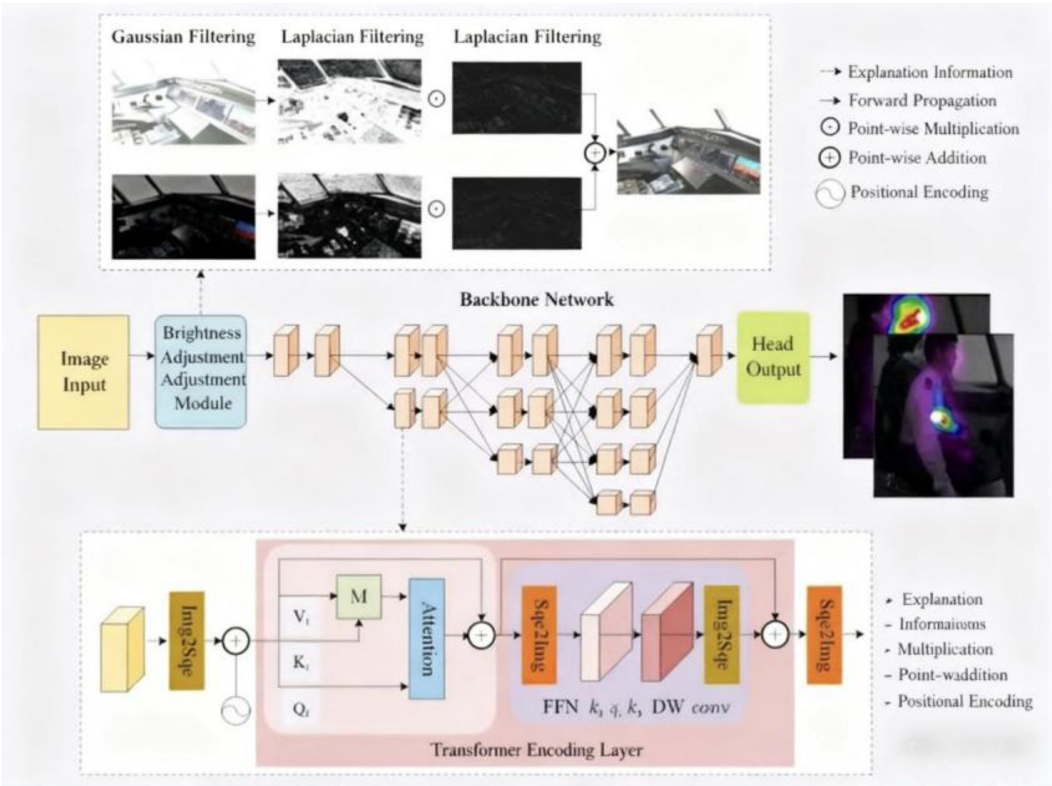

Fig. 2 Overall architecture of pilot limb key point detection in complex lighting environment

3.3 Action and Posture Feature Extraction

3D key points form the foundation of feature extraction. Their coordinates enable quantification of core parameters: For instance, calculating body pitch angle using the head-to-trunk key point line (which helps control pilot scenario errors within ±3° and differentiate between normal sitting posture and lowered-head positions [8]); determining elbow joint angles through the upper-arm forearm key point vector angle to evaluate support operation compliance [9]; and establishing hand-control lever relative positioning via Euclidean distance calculation to provide operational range quantification benchmarks [10].

Dynamic temporal features require integration with image sequence extraction, including operational speed, acceleration, and trajectory curvature (indicating motion smoothness) [11]. The fusion of spatial and temporal features is crucial for enhancing recognition performance: CNN extracts single-frame spatial features while LSTM or Transformer captures temporal dependencies. Their combination forms a “spatiotemporal feature vector”. For instance, in pilot’s coordinated operation recognition between “throttle push-pilot control stick”, the CNN-LSTM model demonstrates higher feature differentiation accuracy compared to standalone CNN [12].

3.4 Action and Posture Recognition Algorithm

3.4.1 Traditional machine learning methods

Support Vector Machines (SVM) and Hidden Markov Models (HMM) are particularly effective for simple motion classifications, such as the binary “push/pull lever” scenario. SVM achieves classification accuracy up to 90% by mapping feature vectors through kernel functions. While HMM models sequential actions using state transition probabilities to identify three states (“push/pull/hold throttle lever”), it maintains an accuracy of approximately 85%, though demonstrating limited adaptability to complex multi-joint coordinated motions. [13, 14]

3.4.2 Deep learning methods

3D CNN directly processes spatiotemporal information in image sequences, achieving 92% accuracy in multi-action recognition during the landing phase of pilots, but suffers from high computational demands with single-frame inference requiring 20ms [15]. The CNN-LSTM fusion model first extracts spatial features through CNN, then applies LSTM to model temporal dependencies, reducing latency and improving accuracy to 89% in joystick operation recognition [15]. Transformer employs self-attention mechanisms to capture long-term action correlations, demonstrating a 5%-8% accuracy improvement over

LSTM in complex operational scenarios like multi-device coordination during landing phases, though it requires larger training datasets [16].

4 Application Scenarios and Typical Research Cases

In flight training evaluation, the multi-view vision system combined with a CNN-LSTM model can identify coordinated control actions of the control stick and throttle during takeoff phase, quantify operational timing errors and amplitude deviations, and assist instructors in precise corrective interventions. The system achieves 91% accuracy rate with 85ms latency, enhancing training efficiency by 15% and driving the development of vectorized precision training.

In cockpit safety monitoring, the 3D CNN-LSTM fusion architecture extracts pilot posture features and analyzes temporal changes. It can identify fatigue-related postures such as prolonged head-down posture and excessive body tilt beyond thresholds, while tracking hand key points to detect operational errors. With a response time of less than 100ms, this system achieved over 200 effective early warnings during transport aircraft testing.[17]

In human-machine interaction optimization for pilots, the dual-camera vision system combined with a CNN-LSTM gesture control model achieves 91.2% accuracy in recognizing gestures like “point, swipe, and click”. By enabling instrument parameter adjustments through gestures, this innovation reduces operational errors during turbulence conditions, enhances operational convenience by 40%, minimizes distraction risks, and optimizes user interaction experience.

5. Conclusions

In summary, multi-view sequence technology has evolved from traditional approaches to deep learning architectures in pilot motion and posture research. Early implementations like SVM and HMM relied on manual features for basic motion classification, establishing technical benchmarks for foundational scenarios. Subsequent advancements—including spatiotemporal modeling of 3D CNNs, real-time optimization of CNN-LSTM systems, and Transformer-based long-sequence collaborative motion capture—have progressively overcome bottlenecks in multi-joint coordination and complex operational process recognition. These innovations now demonstrate practical value in flight training evaluation and cockpit safety monitoring applications.

In the future, balancing data redundancy with lightweight models while integrating physiological signals and other multimodal information to enhance robustness will likely become a critical research direction. Ongoing exploration in this field not only drives advancements in flight safety technologies but also provides a replicable technical framework for motion analysis in complex human-machine interaction scenarios.

References

[1] W.Guan, Z.Luo, W.Jiang, et al. A stereo target pose measurement method based on multi-view vision [J]. Journal of Instrumentation, 2025,46(04):218-227.

[2] Z. Zhang. A Flexible New Technique for Camera Calibration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330-1334.

[3] T. Wang, S. Fu, D. Huang, et al. Pilot Motion Recognition in Cockpit Based on Deep Graphs[J]. Electric Light and Control, 2017, 24(12): 90-94.

[4] X. Zhang, T. Zhu, M. Zhang. Pilot Behavior Recognition Based on Multi-Modality Fusion Technology Using Physiological Characteristics[J]. Journal of Personalized Medicine, 2022, 12(6): 404.

[5] K. Chen, T. Zheng, C. He, et al. A Novel Cross-Resolution Image Alignment Method for Multi-Camera Systems[C]. International Conference on Mobile Computing, Applications and Services. Cham: Springer Nature Switzerland, 2022: 3-14.

[6] H. Liu, Y. Sun, H. Wu, et al. Key Point Detection Method for Civil Aircraft Pilots in Complex Lighting Environments[J/OL]. Journal of Beihang University, 2023, DOI:10.1001-5965.2023.0566 [2025-08-29].

[7] H. Wang, Z. Ji, L. Zhang. Kalman Filter-Based Target Recognition Tracking and Shooting System Design[J]. Journal of Ordnance Equipment Engineering, 2022, 43(11): 286-296.

[8] X. Li, X. Wang. Pilot Posture Quantification Method Based on Multi-Eye Vision[J]. Journal of Aeronautics, 2022, 43(5): 120-130.

[9] J.Wang.Deep High-Resolution Representation Learning for Human Pose Estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI), 2020, 43(10): 3346-3360.

[10] A. Smith. Multi-View 3D Pose Estimation for Pilot Operation Monitoring[C]. IEEE International Conference on Robotics and Automation (IEEE ICRA), 2021: 5678-5684.

[11] X. Zhang, X. Liu. A Review of Time-Sensitive Feature Extraction and Quantization Methods for Human Movement[J]. Journal of Automation, 2023, 49(2): 289-302.

[12] X. Zhao, X. Chen. Flight Operation Sequence Recognition Based on CNN-LSTM Hybrid Model[J]. Journal of Computer Science, 2023, 46(8): 1678-1690.

[13] M. A. Oskoei. Support Vector Machine-Based Classification Scheme for Myoelectric Control Applied to Upper Limb[J]. IEEE Transactions on Biomedical Engineering, 2008, 55(8):

ISSN 2959-6157

1956-1965.
[14] A. F. Bobick. Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(12): 1371-1378.
[15] S.Ji.3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.
[16] H. T. Nguyen. A CNN-LSTM Approach to Human Activity Recognition[J]. IEEE Journal on Biomedical and Health Informatics, 2020, 24(7): 1917-1926.
[17] Y.Wang. Real-Time Fatigue Detection in Cockpits Using 3D CNN-LSTM with Spatiotemporal Attention[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 2567-2578.