Advances in Human Factors of Transportation, Vol. 148, 2024, 12–20
https://doi.org/10.54941/ahfe1005190

Vigilant Air Traffic Control: Gaze-Based Recognition of Detection Failures to Visual Warnings

Zhimin Li and Fan Li

Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, 999077, Hong Kong, China

ABSTRACT

Air traffic controllers sometimes fail to detect visual warnings due to limited attention resources. This challenge would even be exacerbated by the increasing complexity of visual data in future digital tower integrations. Detection failures (DF) manifest in three primary types: ordinary blindness (OB), look but fail to see (LBFTS) error, and misinterpretation (MI), each resulting from disruptions in the detection process stages and necessitating specific countermeasures. This study employs machine learning and eye-tracking in a simulated air traffic control (ATC) environment to identify and differentiate types of DF. Eye movements of 26 participants were tracked across 108 OB, 109 LBFTS, and 95 MI instances to ATC warnings. Seven machine learning models, including three basic and four advanced tree-based models, were assessed for DF recognition. Results found that the gradient boosting decision tree exhibited superior performance with 74% accuracy in four-detection-type recognition, particularly in recognizing OB and LBFTS. Additionally, correct detection and MI are more challenging but still effectively recognized, with correct detection better identified by k-Nearest Neighbour, and MI by light gradient boosting machine. These findings demonstrate the feasibility of real-time gaze-based DF recognition in ATC and offer valuable insights for ATC management in enhancing visual warning detection and aviation safety.

Keywords: Warning detection, Detection failure recognition, Eye-tracking, Air traffic control, Detection failure types

INTRODUCTION

In air traffic control (ATC), automatic warning systems are vital for air traffic controllers (ATCOs) to identify potential conflicts, delivering visual warnings through a human-computer interface (Zhang et al., 2019). However, ATCOs fail to detect visual warnings sometimes due to their limited attention resources (Ruskin et al., 2021). This difficulty would be further amplified by the increased complexity of visual information from the integration of digital towers. A variety of studies indicate that detection failures (DF) could occur in many ways throughout the human monitoring process, which can be summarized into three key types: ordinary blindness (OB), look but fail to see (LBFTS) error, and misinterpretation (MI) (Bruder and Hasse, 2020; Wang et al., 2023; Ruskin et al., 2021). These types are grounded in disruptions to Endsley's widely recognized situation awareness theory, which encompasses perception, comprehension, and projection (Endsley, 1995). Perception refers

© 2024. Published by AHFE Open Access. All rights reserved.

to the detection of stimulus in the environment, comprehension involves synthesizing this information to understand the current situation, and projection is the ability to foresee future states of the environment based on this understanding. These stages collectively contribute to effective detection and decision-making, particularly in dynamic contexts where rapid response to changes is crucial.

The conceptual definition of DF types, as illustrated in Figure 1, highlights how each DF type corresponds to a specific disruption in Endsley's key cognitive stages. Each type has distinct causes and requires specific countermeasures (Causse et al., 2016). Specifically, OB, where warnings are completely overlooked, suggests a need for improved warning displays (Hollnagel, 2000). LBFTS, arising from gaps in cognitively processing perceived warnings, indicates an overload of cognitive capacity and necessitates a reduction in task load (Wang et al., 2022). MI occurs when operators notice but misunderstand warnings and make an erroneous decision, pointing to the need for enhanced personnel training (Bruder and Hasse, 2020). Hence, recognizing DF types to the visual warnings should be valuable for developing efficient interventions accordingly.

Recent studies have shown notable eye movement patterns linked to the DF phenomenon (Li et al., 2023; Mengtao et al., 2023), revealing the potential of eye-tracking-enabled DF recognition. Metrics like fixations, saccades, and pupil responses are identified as reliable indicators of understanding attention shifts and cognitive load (Li et al., 2019). Fixation duration, reflecting cognitive processing depth, along with frequent fixations in relevant areas during attention failures and prolonged gaze on distractors, offer insights into attentional dynamics (Bodala et al., 2017; Bruder and Hasse, 2020). Saccade amplitude relates to scanning strategies (Bodala et al., 2016), while pupil diameter changes suggest cognitive effort variations, especially under DF conditions (Moacdieh et al., 2023). Additionally, first view time analysis in various contexts aids in comprehending visual responses and cognitive workload during DF (Ruscio et al., 2015; Li et al., 2022). However, to what extent can the eye movement features be achieved in recognizing DF types in a safety-critical ATC context remains unexplored.

This study addresses the gap by employing multiple machine learning methods and eye-tracking features for the recognition of DF types in simulated ATC settings. An empirical study was conducted using the 'Endless ATC' simulation platform, involving 26 participants who monitored airspace within an automated warning system. Additionally, we employed and compared seven machine learning methods for DF type recognition, all of which have been extensively adopted in previous literature for their proven effectiveness (Li et al., 2024; Shams et al., 2023). The three basic models, support vector machine (SVM), k-Nearest Neighbour (kNN), and multilayer perceptron (MLP), are effective in high-dimensional spaces and straightforward classification tasks (Cristianini and Shawe-Taylor, 2000). Advanced tree-based ensemble models like random forest, eXtreme Gradient Boosting (XGBoost), gradient boosting decision tree (GBDT), emerging since the early 2000s, are recognized for their robustness and enhanced capabilities in handling complex datasets (Natras et al., 2022). Notably, light gradient boosting machine (LightGBM), introduced by Microsoft in 2017, stands

out for its exceptional efficiency and performance, particularly with large datasets (Ke et al., 2017).

In summary, this study offers a robust scientific approach for key DF-type recognition within ATC warning systems, using eye-tracking data in simulated ATC scenarios. It provides key foundations for ATC management in selecting appropriate countermeasures against DF types, thereby enhancing visual warning detection, and ultimately improving aviation safety.

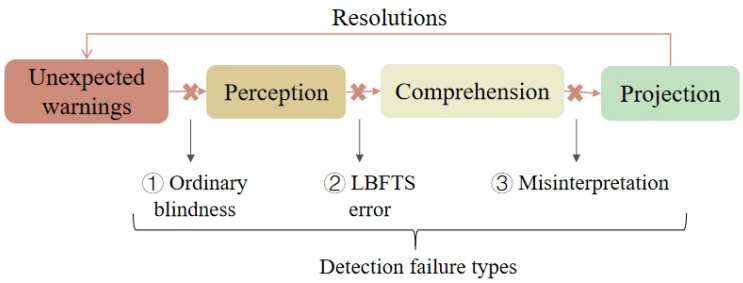


Figure 1: The definitions of DF types to the unexpected warnings.

METHOD

Participants and Apparatus

In a simplified experimental study conducted at The Hong Kong Polytechnic University, we designed a simulation to approximate the general model of ATC monitoring tasks. The participants were university students who possessed normal or corrected-to-normal vision within the age range of 20 to 30 years ($M = 25.65$, $SD = 2.69$), and consented to partake following ethical approval (HSEARS20211117002). The apparatus included a 27-inch monitor paired with a Gazepoint 3 eye tracker to precisely capture eye movement data, essential for understanding the participants' engagement with the task.

Experiment Procedures

All participants were introduced to the Endless ATC platform, a simplified ATC environment. They were tasked with monitoring virtual airspace, identifying aircraft types by their labels, and responding to a set of predefined warnings—each designed to reflect potential ATC alerts.

The experiment proceeded through a structured sequence of phases: an introductory briefing on objectives and methods, a 40-minute training session to familiarize participants with aircraft types and warnings, a practice phase to familiarize themselves with the simulation's interface and tasks, a break and eye tracker calibration, and the formal supervision task with an experimental interface shown in Figure 2. In the task, participants watched a pre-recorded 50-minute video of aircraft control, tracking aircraft and identifying successful takeoffs and landings. They also need to recognize and record eight specific warning types which were displayed for 10 seconds each. The 10-second warning display time was chosen as participants usually noticed and recorded warnings within 8 seconds during practice sessions, making this timeframe sufficient for perception. Overall, each participant encountered a total of 189 warnings. The simulation, while a generalized model of

ATC tasks without the full complexity of actual ATC contexts, was sufficient to induce DF, thus serving the study's aim to understand general monitoring behaviours and responses to potential aerial conflicts.

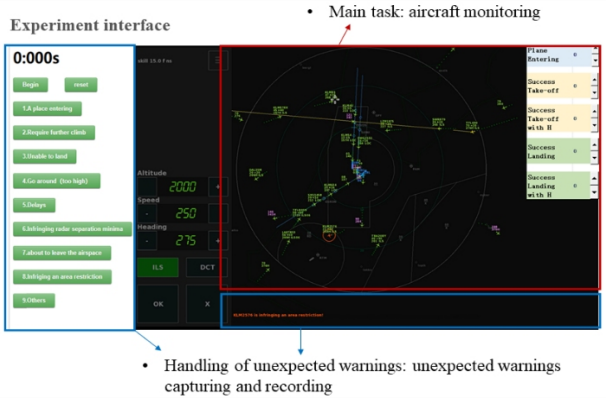


Figure 2: The experimental interface and the tasks of participants.

DF Type and Gaze Feature Measurement

In the experiment, detection failures were categorized based on subjects' gaze and response to warnings. OB was identified when there was no fixation on the warnings, indicating a complete miss. LBFTS occurred when subjects fixated on the warnings but failed to record them. MI errors were noted when subjects fixated on and recorded the warnings, but the recording was incorrect. Correction detection (CD) was achieved when subjects fixated on the warnings and recorded them accurately. Eye movement data were collected for each of the above four detection types, as shown in Figure 3. In cases of CD and MI, where subjects perceived and recorded the warnings, eye movements were tracked from the warning's appearance to its recording. For OB and LBFTS instances, where warnings were either perceived or responded to, eye movements were collected for the entire 10-second warning display duration.

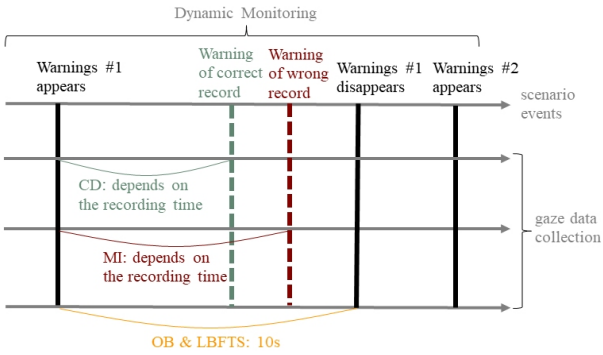


Figure 3: The eye movement data collection of different DF types during the experiment.

This study focused on analysing eye movement patterns in response to various detection types. Key variables included the first view time (FVT) in seconds, fixation count (FC) and duration (FD) within the area of interest (AOI), i.e. warning display area, blink frequency (BF), mean saccade amplitude (MSA), and mean pupil diameter (MPD) in response to warnings. MPD was further broken down into mean left (MLPD) and right (MRPD) pupil diameters. Specifically, the first view time for the OB type is standardized at 10 seconds, reflecting their non-responsiveness to warnings. In contrast, the first view time for the other three detection types is determined by the time from the onset of the warning to the first viewing. To normalize individual differences in pupil diameter and blink frequency in the collected eye movement data, we used baseline normalization. The baseline for pupil diameter and blink frequency was set during the thirty seconds following the first five minutes of the experiment. Task-evoked pupillary responses were determined by comparing the MPD and BF at warning onset with this baseline.

Data Analysis

In the data analysis, we employed seven regression models for DF type recognition, each with unique strengths. The basic models included SVM, known for handling high-dimensional spaces; kNN, effective in pattern recognition due to its simplicity; and MLP, a neural network adept at learning nonlinear relationships. Additionally, we utilized advanced tree-based ensemble models: random forest, which enhances accuracy through multiple decision trees; XGBoost, recognized for its speed and efficiency in structured data; GBDT, focusing on correcting previous tree errors sequentially; and LightGBM, optimized for large data sets and high-speed processing. These models were chosen for their proven classification capabilities and were systematically compared to determine the optimal approach for DF type recognition in our context. The input features and outputs are the seven eye movement features and the classification of four detection types, respectively. In assessing our models, we used four metrics: accuracy, precision, recall, and F1 score. Besides, we applied 5-fold cross-validation to enhance the reliability of our results, reducing overfitting by training and testing on varied data segments.

RESULTS

The study recorded eye movements during 108 OB, 109 LBFTS, and 95 MI instances to ATC warnings. The performance analysis of seven machine learning models for classifying eye-tracking data into four detection types is shown in Table 1. The table demonstrates GBDT emerges as the top-performing model with uniform scores of 0.74 in accuracy, precision, recall, and F1 score, reflecting a high level of four-type classification efficacy. Random Forest and LightGBM models also exhibit substantial efficacy, particularly Random Forest, which parallels GBDT in accuracy, precision, and recall. Among the simpler basic models, MLP surpass SVM and kNN, the latter trailing with metrics around 0.65. This analysis underscores the advanced models' enhanced ability to navigate the intricacies of multi-class eye-tracking data classification.

Following the global results assessment, a comparative analysis of model performances across detection categories (i.e. CD, OB, LBFTS, and MI) is conducted, as shown in Figure 4. Higher prediction performance in classifying a detection type indicates more distinguishable eye movement features, revealing how models uniquely respond to each type's distinctive eye-tracking features. Figure 4 reveals a marked variance in model efficacy across these detection categories. Notably, the models exhibit a varied degree of proficiency across the four detection types. OB stands out with the highest prediction accuracy, followed by LBFTS. Conversely, the assessment of CD and MI presents more complex scenarios. A detailed analysis of the prediction outcomes for each category will be conducted sequentially.

Table 1. Performance of the compared models for classifying eye-tracking data into four detection types.

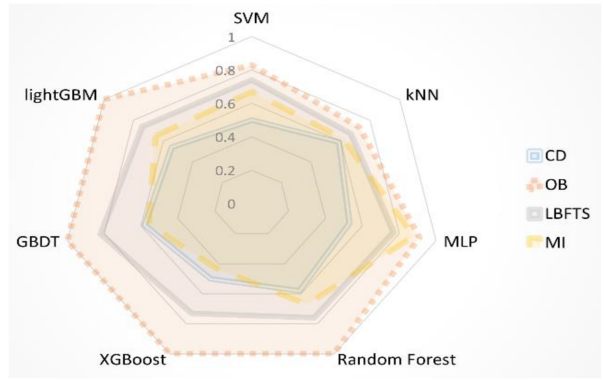
	SVM	kNN	MLP	Random Forest	XGBoost	GBDT	LightGBM
Accuracy	0.67	0.65	0.69	**0.74**	0.67	**0.74**	0.73
Precision	0.68	0.65	0.69	**0.74**	0.66	**0.74**	0.73
Recall	0.67	0.66	0.69	**0.74**	0.67	**0.74**	0.73
F1 score	0.66	0.64	0.67	0.73	0.66	**0.74**	0.73

In the OB category, characterized by an absence of fixation on warnings, models, especially four advanced tree-based models, showcase exceptional accuracy, with precision and recall scores reaching up to 1.00. This remarkable performance is indicative of the distinctive eye movement patterns in OB, where the lack of fixations and prolonged first view time offer clear markers for classification (Hergovich and Oberfichtner, 2016).

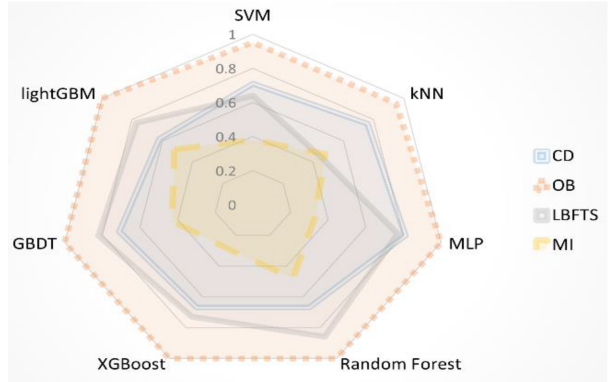
In LBFTS, distinct eye movement patterns emerge, despite its similarities with CD and MI regarding fixations on warnings. The performance of GBDT and Random Forest models in LBFTS is noteworthy: GBDT achieves a precision, recall, and F1 score of 0.82, while Random Forest records a precision of 0.76, recall of 0.86, and F1 score of 0.81. These results highlight the unique eye movement patterns in LBFTS, contrasting with the focused attention in CD and the attentional inaccuracies in MI. This distinction is consistent with previous research indicating inattentional blindness correlates with fewer fixations on stimuli, increased pupillary diameter, and less efficient attention allocation (Moacdieh et al., 2023; Richards et al., 2012). The results highlight the significance of eye movements as key indicators for recognizing LBFTS, demonstrating their pivotal role in DF classification.

The MI and CD categories, characterized by their respective incorrect and correct recordings of warnings, pose distinct analytical challenges. Comparatively, MI demonstrates superior predictive outcomes over CD, exemplified by the MLP model's prediction score of 0.88 for MI, surpassing CD's highest score of 0.59. However, in terms of recall metric, CD outperforms MI; the MLP model achieves a recall of 0.81 for CD, whereas MI's highest recall is 0.52, achieved by the Random Forest model. A comprehensive evaluation of the models' effectiveness, as assessed through the F1 score, reveals

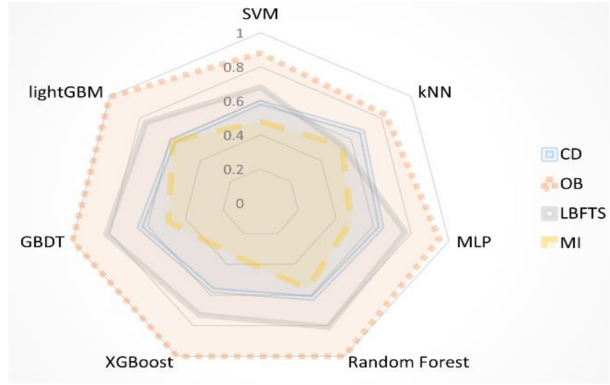
that CD's seven model outcomes are more robust than those for MI. Specifically, the kNN model for CD attains the highest F1 score of 0.67, compared to MI's peak score of 0.58 by the lightGBM model. These findings suggest that CDs are generally more discernible than MIs, supporting the feasibility of distinguishing them through eye movement analysis, despite inherent challenges.



(a) Precision



(b) Recall



(c) F1 score

Figure 4: A comparative analysis of model performances across four detection categories.

CONCLUSION

In summary, this study effectively employed machine learning models to identify three DF types in response to warnings, utilizing eye movement data gathered from a specifically simulated ATC context. In a global analysis of machine learning models, gradient boosting decision tree stand out for their effectiveness in classifying four detection types. This demonstrated the feasibility for real-time gaze-based recognition systems in ATC, particularly in promptly identifying instances where warnings are missed (OB) or observed but not acted upon (LBFTS). In contrast, CD and MI categories are more challenging to classify, with less but still effective prediction accuracy. The k-Nearest Neighbour model demonstrates heightened proficiency in recognizing CD, while light gradient boosting machine perform better in identifying MI. The study's approach and findings have significant implications for enhancing visual warning detection and improving human-computer interaction in aviation, ultimately contributing to safer and more efficient ATC operations.

The study's limitations include a small participant pool and the use of a generalized supervision simulation rather than an authentic ATC environment, which may limit the findings' applicability. However, this preliminary work lays the groundwork for more realistic scenario-based experiments in the future. Focusing mainly on eye-tracking metrics, further research could benefit from incorporating broader physiological data like electroencephalogram.

ACKNOWLEDGMENT

This work was supported by the Hong Kong Polytechnic University under Grant P0038827 and Grant P0038933. This study has been granted human ethics approval from the PolyU Institutional Review Board of The Hong Kong Polytechnic University (IRB Reference Number: HSEARS20211117002).

REFERENCES

Bodala I P, Abbasi N I, Sun Y, et al. (2017) of Conference. Measuring vigilance decrement using computer vision assisted eye tracking in dynamic naturalistic environments [C], IEEE; City. 2478–2481.

Bodala I P, Li J, Thakor N V, et al. (2016). EEG and eye tracking demonstrate vigilance enhancement with challenge integration [J]. Frontiers in human neuroscience, 10: 273.

Bruder C, Hasse C (2020). What the eyes reveal: Investigating the detection of automation failures [J]. Appl Ergon, 82: 102967.

Causse M, Imbert J-P, Giraudet L, et al. (2016). The role of cognitive and perceptual loads in inattentional deafness [J]. Frontiers in human neuroscience, 10: 344.

Cristianini N, Shawe-Taylor J (2000). An introduction to support vector machines and other kernel-based learning methods [M]. Cambridge university press.

Endsley M R (1995). Toward a theory of situation awareness in dynamic systems [J]. Human factors, 37: 32–64.

Hergovich A, Oberfichtner B (2016). Magic and misdirection: The influence of social cues on the allocation of visual attention while watching a cups-and-balls routine [J]. Frontiers in Psychology, 7: 761.
Hollnagel E (2000). Looking for errors of omission and commission or The Hunting of the Snark revisited [J]. Reliability Engineering & System Safety, 68: 135–145.
Ke G, Meng Q, Finley T, et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree [J]. Advances in neural information processing systems, 30.
Li F, Chen C-H, Lee C-H, et al. (2022). Artificial intelligence-enabled non-intrusive vigilance assessment approach to reducing traffic controller's human errors [J]. Knowledge-Based Systems, 239: 108047.
Li F, Lee C-H, Chen C-H, et al. (2019). Hybrid data-driven vigilance model in traffic control center using eye-tracking data and context data [J]. Advanced Engineering Informatics, 42: 100940.
Li Z, Li R, Yuan L, et al. (2024). A benchmarking framework for eye-tracking-based vigilance prediction of vessel traffic controllers [J]. Engineering Applications of Artificial Intelligence, 129: 107660.
Li Z, Li Z, Li F (2023). Visual Attention Analytics for Individual Perception Differences and Task Load-Induced Inattentional Blindness [C], Springer; City. 71–83.
Mengtao L, Fan L, Gangyan X, et al. (2023). Leveraging eye-tracking technologies to promote aviation safety-a review of key aspects, challenges, and future perspectives [J]. Safety Science, 168: 106295.
Moacdieh N M, Dibo M, Halabi Z, et al. (2023) of Conference. Eye tracking to evaluate the effectiveness of electronic medical record training [C]; City. 1–7.
Natras R, Soja B, Schmidt M (2022). Ensemble machine learning of Random Forest, AdaBoost and XGBoost for vertical total electron content forecasting [J]. Remote Sensing, 14: 3547.
Richards A, Hannon E M, Vitkovitch M (2012). Distracted by distractors: Eye movements in a dynamic inattentional blindness task [J]. Consciousness and Cognition, 21: 170–176.
Ruscio D, Ciceri M R, Biassoni F (2015). How does a collision warning system shape driver's brake response time? The influence of expectancy and automation complacency on real-life emergency braking [J]. Accident Analysis & Prevention, 77: 72–81.
Ruskin K J, Corvin C, Rice S, et al. (2021). Alarms, alerts, and warnings in air traffic control: An analysis of reports from the Aviation Safety Reporting System [J]. Transportation research interdisciplinary perspectives, 12: 100502.
Shams M Y, Elshewey A M, El-Kenawy E-S M, et al. (2023). Water quality prediction using machine learning models based on grid search method [J]. Multimedia Tools and Applications: 1–28.
Wang S, Liu Y, Li S, et al. (2023). The effect of two-stage warning system on human performance along with different takeover strategies [J]. International Journal of Industrial Ergonomics, 97: 103492.
Wang Y, Wu Y, Chen C, et al. (2022). Inattentional blindness in augmented reality head-up display-assisted driving [J]. International Journal of Human–Computer Interaction, 38: 837–850.
Zhang J, Liu H, Yang Q (2019) of Conference. Design and Implementation of Runway Incursion Automatic Warning Simulation System [C], IEEE; City. 611–613.