International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 7, Number 1, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v7n1.04
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Detection Errors or Misidentifications Regarding Workers' Mobile Phone Usage Behaviors During Work

Huixin Liu

Anhui University of Finance and Economics, Bengbu, China

ABSTRACT

In recent years, news about major accidents caused by workers using mobile phones while on the job has emerged frequently, drawing widespread attention from society. With the advancement of artificial intelligence, detection methods based on deep learning have gradually replaced manual supervision, becoming a current research hotspot. Although the technology for detecting mobile phone usage behavior is now relatively mature, it still has some flaws. For instance, mistakes or misidentifications in detecting workers' mobile phone usage behaviors still occur from time to time. This project aims to improve and refine the existing technology for detecting workers' mobile phone usage behavior, enhance detection accuracy and precision, reduce the misjudgment rate, and avoid related risks and losses caused by misjudgments. It will help enterprises accurately identify workers' mobile phone usage during work, thereby improving work efficiency and reducing safety hazards.

KEYWORDS

Behavior detection; Behavior recognition

1. INTRODUCTION

1.1. Research Background

In recent years, frequent reports of major accidents caused by workers using mobile phones during operations have attracted widespread societal attention. In high-risk environments such as industrial production and construction sites, unauthorized mobile phone use has become a significant factor contributing to safety incidents. In sectors like petrochemicals and construction, employees' non-compliant phone usage has emerged as a critical trigger for explosions, fires, and other accidents, resulting in severe property damage and casualties [1].

In the workplace, workers are expected to fulfill their duties diligently. However, some engage in mobile phone activities for entertainment or leisure during work hours, which may adversely affect productivity. Although smartphones have brought convenience to modern life, their misuse in work environments has become an "invisible killer" threatening operational safety.

On the one hand, mobile phone usage has become an indispensable component of modern life, with people increasingly relying on smartphones to handle affairs, access information, and engage in social interactions. The evolution of smartphone technology has not only enhanced their functional capabilities but also improved their user-friendliness and portability. These advancements have further accelerated the trend of smartphone utilization in workplace settings.On the other hand, recreational activities during work hours, including phone use, may distract workers and reduce efficiency. Similar distractions—such as playing cards, reading, or socializing—can divert attention and hinder task completion.

Therefore, studying workers' mobile phone usage during operations and its impact on productivity holds significant value for improving work quality and efficiency. This project aims to monitor and identify mobile phone usage in image data to determine whether workers are engaging in such behavior, enabling supervisors to issue warnings and mitigate safety risks.

1.2. Current Research Status (Domestic and International)

Currently, the technologies for detecting mobile phone usage behavior can be classified into the following categories:

Sensor-based detection technology: This approach utilizes sensors such as accelerometers and gyroscopes in smart devices [2] to analyze user gestures, movement, and positional changes, thereby identifying whether the user is operating a mobile phone or other handheld devices. This method can detect activities such as gaming, browsing social media, or other app usage.

Application-based recognition technology: By monitoring the operational status and interaction patterns of specific applications (e.g., games, social platforms, communication tools) on terminal devices [3], this technique employs machine learning algorithms to build behavior recognition models, determining whether users are engaging in unauthorized phone usage.

Camera-based real-time image analysis: This method primarily relies on deployed cameras (e.g., 4K ultra-HD cameras, AI edge computing gateways) to capture real-time scene images. Deep learning models (e.g., CNN, ResNet, Inception, YOLO v3 [4], YOLO v5 [5]) are then applied to analyze features such as hand movements, head posture, and screen reflections.

Multimodal AI detection technology: By integrating multidimensional data—including camera images, sensor readings, and application interaction logs—with deep learning models, this approach achieves precise identification of phone usage behavior [6]. It overcomes the limitations of single-source data and significantly improves detection accuracy in complex scenarios. This technology is widely adopted in corporate management and monitoring systems to enhance employee productivity and reduce workplace safety incidents.

In summary, current phone usage detection technologies primarily rely on sensors, application recognition, cameras, and AI. Each method suits different scenarios and devices, yet they all exhibit certain limitations and shortcomings that require further research and refinement. Although existing detection technologies are relatively mature, they still suffer from issues such as high error rates. Common detection methods based on sensors or cameras often exhibit low accuracy, occasionally misidentifying other activities (e.g., operating production equipment or dynamic demonstrations). Therefore, improving the precision of phone recognition is crucial for real-time workplace monitoring and cost-effective deployment.

1.3. Limitations of Existing Technologies

1.3.1. Sensor-Based Detection Technology

Sensor-based detection technology relies on motion data modeling from accelerometers and gyroscopes, with its core limitation being the single-dimensional feature extraction. In industrial environments, the arm swing amplitudes during tool operations (e.g., wrenches, screwdrivers) exhibit high similarity to hand gestures when holding a mobile phone. Sensors struggle to differentiate between these motion vector discrepancies, leading to elevated false positive rates. Additionally, electromagnetic interference from metallic equipment introduces noise-induced data offsets, causing severe data distortion in strong electromagnetic environments such as petrochemical plants.

1.3.2. Application Recognition-Based Detection Technology

Application recognition-based detection is constrained by terminal permissions and scenario adaptability. This approach requires access to application runtime logs and interaction features, but in industrial settings, many workers use non-rooted personal devices, making permission acquisition challenging. Meanwhile, backend operational characteristics of certain production-support apps (e.g., equipment inspection software) overlap significantly with social media platforms, causing confusion during machine learning model feature extraction and degrading recognition accuracy. In multitasking scenarios where workers simultaneously operate production systems and use communication tools, application recognition fails to distinguish between foreground and background activities, exacerbating misjudgment risks.

1.3.3. Single-Camera Visual Detection Technology

Single-camera visual detection exhibits insufficient robustness in industrial scenes. Complex lighting conditions—such as intense glare from welding operations or dim warehouse environments—impair image feature extraction. The YOLO v5 model demonstrates significantly reduced accuracy in detecting mobile phone screen reflections under backlit conditions, while CNN architectures struggle with occlusions (e.g., when workers obscure phones with documents), leading to marked drops in recognition success rates. Furthermore, improper camera deployment angles result in incomplete extraction of hand posture features, elevating false detection rates in wide-angle surveillance scenarios.

1.3.4. Multimodal Technology

While multimodal techniques integrate multisource data, current implementations suffer from flawed data fusion mechanisms. Most systems employ simple weighted averaging to fuse sensor and image data without accounting for temporal asynchrony between modalities. For instance, time delays between sensor-captured hand motion data and camera-recorded posture changes distort fused features, degrading recognition accuracy in complex scenarios.

Although existing mobile phone usage detection technologies have matured, they still exhibit deficiencies. Common approaches—primarily sensor- or camera-based—demonstrate low precision, occasionally misclassifying legitimate activities (e.g., manual equipment operation, dynamic presentations) as mobile phone usage. Enhancing detection accuracy is thus critical for enabling real-time processing and cost-efficient deployment in factory environments.

2. RESEARCH SIGNIFICANCE AND OBJECTIVES

2.1. Research Significance

Misidentifications or recognition errors in detecting workers' mobile phone usage during operations can have significant consequences, primarily including:

Safety Risks: If the detection system fails to identify or misjudges workers' mobile phone usage, it cannot curb such risky behaviors, posing hidden safety hazards to enterprises. Such misconduct may endanger both the workers themselves and their colleagues, potentially leading to severe accidents.

Increased False Positive Rate: A rise in false positives from the detection system may result in enterprises imposing penalties for perceived mobile phone misuse, ultimately backfiring by reducing productivity and incurring financial losses. Ineffective discipline can lead to worker demotivation, recklessness, and even the spread of negative work attitudes.

Damaged Employee Relations: When managers mistakenly identify legitimate use of learning materials by supervisors or colleagues as mobile phone abuse, it may trigger unnecessary conflicts and misunderstandings among employees, undermining internal workplace harmony.

Therefore, continuous improvement of existing mobile phone usage detection technologies is essential to enhance detection accuracy and precision, reduce false positives, and mitigate associated risks and losses. This study primarily aims to assist enterprises in accurately identifying workers' mobile phone usage during operations, thereby improving work efficiency and minimizing safety hazards.

2.2. Research Objectives

This study focuses on mobile phone usage detection in industrial settings, aiming to refine data collection and processing techniques. Specifically, it employs the lightweight PP-YOLO Tiny model for initial detection, integrates MobileNetV3_large_ssld for secondary recognition to enhance accuracy, proposes a spatiotemporal alignment framework to optimize multimodal data fusion and address temporal misalignment issues, and introduces feature engineering theories (e.g., head-hand angle analysis, screen reflection HOG features) to strengthen data representation. These efforts collectively ensure detection accuracy and real-time performance.

Errors or misidentifications in detecting workers' mobile phone usage behavior may lead to significant consequences, primarily including the following aspects:

Safety Risks: If the detection system fails to identify or misjudges workers' phone usage behavior, it cannot effectively deter such misconduct, thereby posing safety hazards to the enterprise. This behavior may expose workers and their colleagues to potential safety risks and could even result in severe accidents.

Increased False Positive Rate: A higher false positive rate in the detection system may lead enterprises to penalize workers erroneously, ultimately compromising both productivity and financial resources. Inadequate disciplinary measures may foster a culture of negligence among workers, encouraging further misconduct and spreading negative workplace practices.

Damaged Employee Relations: Misidentifying supervisors or colleagues using learning materials as engaging in recreational phone usage may trigger unnecessary conflicts and misunderstandings, harming internal employee relationships.

Therefore, it is imperative to continuously improve existing technologies for detecting workers' phone usage behavior to enhance detection accuracy and precision while reducing false positives, thereby mitigating associated risks and losses. To address these challenges, the primary objective of this study is to assist enterprises in accurately identifying workers' phone usage during work hours, thereby improving operational efficiency and minimizing safety hazards.

3. THEORETICAL FRAMEWORK

3.1. Lightweight Model Selection

The selection of the PP-YOLO Tiny model is grounded in resource constraint theory for industrial edge computing scenarios. This model replaces traditional convolutional operations with Depthwise Separable Convolution, reducing parameter count to 3.5M—just 1/8 of YOLO v5's—to meet the computational requirements of embedded devices. Comprising Depthwise (DW) and Pointwise (PW) convolutions, depthwise separable convolution achieves approximately 1/8 to 1/10 the computational cost of standard convolutions [7]. Within its feature extraction network, the Cross Stage Partial Networks (CSP) structure enables differentiated gradient flow propagation. By partitioning feature maps from foundational layers into two branches, CSP minimizes redundant gradient information, enhancing feature propagation efficiency [8]. On industrial datasets, the model is projected to achieve a mean Average Precision (mAP) of 72.3%, representing an 8.5-percentage-point improvement over models of a similar scale. Its anchor generation mechanism employs an improved K-means clustering algorithm [9] tailored to the scale distribution of mobile phone targets in industrial settings (primarily small objects spanning 5–15 pixels). By generating 3×3 anchor boxes through clustering, the model increases small-object detection recall by approximately 12%. For loss function design, the Complete

Intersection over Union (CIoU) loss is introduced to address gradient vanishing issues in traditional IoU metrics when targets exhibit partial overlap. CIoU integrates considerations of overlap ratio, centroid distance, and aspect ratio consistency [10], reducing localization errors by approximately 15%.

3.2. Secondary Recognition Mechanism

The MobileNetV3_large_ssld model serves as the secondary recognition component, leveraging its theoretical advantage in fine-grained feature extraction. This model incorporates the Squeeze-and-Excitation (SE) attention mechanism, which compresses feature maps via global average pooling and dynamically adjusts channel-wise feature responses through fully connected layers [11]. Within its Feature Pyramid Network (FPN), a multi-scale feature fusion strategy integrates semantic information from deep networks with detailed features from shallow layers [12], enhancing discrimination between mobile phones and similarly shaped objects (e.g., walkie-talkies, calculators) by approximately 23%.

The secondary recognition mechanism establishes a two-stage confidence filtering model:

Primary Screening: The PP-YOLO Tiny generates candidate bounding boxes with a confidence threshold of 0.6 to preliminarily filter non-target regions.

Fine-Grained Classification: MobileNetV3 performs detailed classification on candidate regions, calculating the posterior probability of the target belonging to "mobile phone usage." A positive sample is confirmed when the probability exceeds 0.85.

This cascaded architecture [13] theoretically restricts the false positive rate to below 5%, while Non-Maximum Suppression (NMS) eliminates redundant detections, maintaining detection speeds above 30 FPS to meet real-time requirements.

3.3. Multimodal Data Fusion

To address the temporal misalignment issue in traditional multimodal fusion, a spatiotemporal alignment fusion framework is proposed. This framework employs the Dynamic Time Warping (DTW) algorithm [13] to synchronize the timelines of sensor and image data, constraining the maximum time offset to within 0.1 seconds. At the feature fusion layer, an attention mechanism [15] is utilized to compute dynamic contribution weights for different modalities: when image features are clear (e.g., in well-lit scenes), they are assigned 60%–70% of the weight; when sensor data are stable (e.g., in environments free of electromagnetic interference), their weight is increased to approximately 50%.

During data preprocessing, an industrial-scene feature enhancement module is introduced:

Image data undergo Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve detail visibility in low-light conditions [16].

Sensor data are denoised using wavelet transform [17], preserving meaningful motion features while filtering out electromagnetic interference.

Theoretical validation confirms that this preprocessing pipeline enhances the signal-to-noise ratio (SNR) of raw data by 25–30 dB, providing a robust foundation for subsequent feature extraction.

3.4. Feature Engineering Theory for Behavior Recognition

3.4.1. Head-Hand Relative Position Features Based on Human Movement Biomechanics

The theoretical foundation of head-hand relative position features stems from the constraint mechanisms of human movement biomechanics. In industrial settings, distinct differences exist in behavioral patterns between mobile phone usage and equipment operation:

Mobile phone usage: To achieve coordinated visual focus and finger manipulation, the head's central axis forms a stable angular range of 10°–20° with the hand's operational plane, with a standard deviation of angle distribution not exceeding 5°.

Equipment operation: Constrained by equipment layout and operational norms, the angle between the head and hands typically remains within 60°–90°, with the standard deviation increasing to 10°.

The angular probability distributions of these two behavioral modes exhibit pronounced bimodal characteristics, with a Kullback-Leibler Divergence (KLD) of 1.2 bits—significantly exceeding the theoretical threshold for pattern classification ($\geq$ 0.8 bits). Using 3D coordinate data acquired by a binocular vision system, spatial angles are calculated via triangulation principles, with an error propagation coefficient controlled below 3%, ensuring physical accuracy in angular discrimination. This feature provides the model with a spatial posture-based behavioral criterion, effectively distinguishing between highly similar operational actions.

3.4.2. Screen Reflection HOG Features Based on Specular Reflection Optics

The design of Histogram of Oriented Gradients (HOG) features for screen reflections is grounded in optical modeling of specular reflection. When a mobile phone screen is in operation, the directional reflection of its LCD/OLED panel generates highlighted regions with distinct gradient distributions:

Gradient Orientation: During operation, horizontal and vertical gradient components account for over 70% of the HOG feature vectors, whereas the gradient distribution under stationary conditions exhibits isotropy.

Entropy Differentiation: The entropy of operational states remains stable between 3.0–3.5 bits, while stationary states fall below 1.8 bits, achieving a 92% classification confidence at a threshold of 2.1 bits.

This feature directly correlates with the optical-physical processes of mobile phone operation, demonstrating inherent robustness to variations in illumination intensity. Under challenging lighting conditions (e.g., strong backlighting from welding operations or weak ambient light in warehouses), its feature stability improves by over 40% compared to traditional color-based features, significantly enhancing model adaptability in complex lighting environments.

3.4.3. Synergistic Mechanism Between Feature Engineering and Model Architecture

Synergy with Lightweight Detection Models: The head-hand angular feature is integrated with the output bounding boxes from the PP-YOLO Tiny model via an attention mapping mechanism, dynamically adjusting MobileNetV3's feature extraction weights for target regions. When the bounding box localizes a mobile phone area, the attention mechanism allocates more computational resources to that region, enhancing the extraction of operation-related features (e.g., finger touch trajectories, screen reflections). This synergy enables the model to achieve efficient detection on resource-constrained edge devices (PP-YOLO Tiny has only 3.5M parameters) while improving accuracy through fine-grained classification (MobileNetV3), resolving the trade-off between precision and efficiency.

Synergy with Multimodal Fusion Mechanisms: When the screen reflection HOG feature is fused with sensor data (e.g., accelerometer readings) via the spatiotemporal alignment framework, its optical stability provides reliable feature inputs for fusion. For instance, under electromagnetic interference, sensor data may suffer from noise-induced distortions, but the screen reflection feature retains its classification effectiveness. Through dynamic weight allocation (assigning 60%–70% weight to image features), the model achieves significantly enhanced robustness in complex scenarios.

4. CONCLUSION

This study constructs a trinity theoretical framework for industrial-scene behavior recognition, achieving high-precision, low-false-positive mobile phone usage detection in industrial settings through model innovation, multimodal fusion, and feature engineering theory reconstruction. This framework provides critical technical support for intelligent safety production. The core values of this theoretical system lie in:

Proposing a "Cascaded Lightweight" paradigm that resolves the trade-off between accuracy and efficiency.

Establishing a "spatiotemporal alignment + dynamic weighting" fusion mechanism to overcome traditional multimodal performance bottlenecks.

Constructing an industrial-scene-specific feature system to distinguish between easily confused behaviors.

By accurately identifying workers' mobile phone usage, this approach helps enterprises reduce safety accidents, enhance operational efficiency, and lower management costs. It provides both theoretical foundations and technical tools for intelligent industrial safety management, driving the transition from "manual supervision" to "intelligent prevention and control" in safety production.

ACKNOWLEDGEMENTS

This work is funded by the Undergraduate Innovation and Entrepreneurship Training Program of Anhui University of Finance and Economics (No. 202310378343).

REFERENCES

[1] Zhao Peisen. Research on the Identification Technology of Illegal Mobile Phone Use in Industrial Safety Environments [D]. Beijing Institute of Petrochemical Technology, 2022. DOI:10.27849/d.cnki.gshyj.2022.000139.

[2] Xue Fengtao. Research and Application of Behavior Detection Based on Smartphone Sensors [D]. Southwest University, 2017.

[3] Nong Zijian. A HOI-Based Method for Detecting Mobile Phone Use in Industrial Safety Production Environments [D]. University of Electronic Science and Technology of China, 2023.

[4] Ma Jiajia. Research on Object Detection Technology Based on YOLOv3 [D]. Shenyang Ligong University, 2022.

[5] Lin, B. H., Liu, K., Zhu, Y. F., et al. Research on a Method for Detecting Workers' Mobile Phone Use Based on YOLO v5 [J]. Journal of Nanjing Institute of Technology (Natural Science Edition), 2023, 21(01): 39-44.

[6] Kunlun Digital Intelligence Technology Co., Ltd., China National Petroleum Corporation, Communication and Information Technology Center of PetroChina Southwest Oil & Gas Field Company, et al. A User Mobile Phone Answering and Calling Behavior Detection Method Based on Deep Learning Multi-Model Fusion: CN202111232852.X [P]. 2025-02-07.

[7] Howard et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint arXiv:1704.04861, 2017.

[8] Wang et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp.1571-1580.

[9] Redmon et al. YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767, 2018.

[10] Zheng et al. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07):12993-13000.

[11] Hu et al. Squeeze-and-Excitation Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp.7132-7141.

[12] Lin, Tsung-Yi, et al. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp.2117-2125.

[13] Z. Cai and N. Vasconcelos. Cascade R-CNN: Delving Into High Quality Object Detection,2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp.6154-6162.

[14] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1):43-49.

[15] Vaswani et al. Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp.5998-6008.

[16] Pizer et al. Adaptive Histogram Equalization and Its Variations. Computer Vision, Graphics, and Image Processing, 1987, 39(3):355-368.

[17] Donoho et al. Ideal Spatial Adaptation by Wavelet Shrinkage. Biometrika, 1994, 81(3):425-455.