Sensors and Materials, Vol. 36, No. 12 (2024) 5247–5266
MYU Tokyo

S & M 3865

Application of 3D Convolutional Neural Networks for Continuous Motion Identification and Behavioral Safety Analysis of Factory Roll Cutting Machine Operators

Chien-Lin Chiang,[1] I-Long Lin,[1] Ming-Yuan Peng,[2]
Wen-Hsin Liang,[3] and Yi-Yuan Chiang[3*]

[1]Department of Computer Science and Engineering, Tatung University,
No. 40, Sec. 3, Zhongshan N. Rd., Taipei City 104, Taiwan (R.O.C)
[2]Graduate Institute of National Development and Mainland China Studies Chinese Culture University,
Full 55, Hwa-Kang Rd., Yang-Ming-Shan, Taipei, Taiwan 11114, R.O.C.
[3]Department of Computer Science and Information Engineering, Vanung University,
No.1, Wanneng Rd., Zhongli Dist., Taoyuan City, Taiwan 320676, R.O.C.

(Received April 30, 2024; accepted December 10, 2024)

Keywords: machine learning, 3D convolutional neural networks, factory safety monitoring, image processing techniques

In this study, we aim to develop an innovative factory safety monitoring system that integrates 3D convolutional neural networks (CNNs) from the field of machine learning with IoT technology to address the safety and efficiency challenges faced by the industrial production sector in the post-pandemic era. By utilizing an improved 3D CNN model, the system can capture and analyze potential safety risks in the factory environment in real time from dynamic video footage, enhancing the understanding and predictive capability of continuous actions and behavior patterns. We demonstrate that 3D CNNs exhibit higher accuracy and dynamic scene analysis capabilities in capturing spatio-temporal features than traditional 2D CNNs. Furthermore, the integration of IoT technology facilitates more efficient data collection, transmission, and real-time analysis, thereby strengthening real-time safety monitoring and decision support. The application and validation of the system in real factory environments have proven its effectiveness in enhancing production line safety and operational efficiency, offering a new solution for the fields of industrial automation and intelligent manufacturing to bolster real-time safety monitoring and decision support.

1. Introduction

Since the outbreak of COVID-19 in early 2020, the pandemic has spread to more than 130 countries and regions within a few months. COVID-19, owing to its high transmissibility, has not only caused severe harm to human life and health but also led to a lifestyle of mutual isolation. This has inflicted unprecedented severe impacts and downturns on the global manufacturing and service industries, which heavily rely on a large workforce for production

*Corresponding author: e-mail: ychiang@mail.vnu.edu.tw
https://doi.org/10.18494/SAM5121

ISSN 0914-4935

and services, leading to disruptions in various industries' supply chains and causing incalculable economic losses worldwide.(1) During the pandemic, companies around the world introduced new remote work models to mitigate the impact on product manufacturing.(2) For the manufacturing industry, remote work inevitably resulted in a reduction in factory labor. However, operations, maintenance, and production management still depend on human resources. As a result, the manufacturing sector has invested more in technological innovation, using smart automation technologies to replace traditional labor, minimizing staffing on production lines to address labor shortages, and mitigating the significant impact of the pandemic on the industry. In the post-pandemic era, global manufacturing and service industries have not only continued the technological innovations adopted during the pandemic but are also facing unprecedented challenges and opportunities to improve manufacturing safety management and production efficiency. In this context, the industrial sector urgently needs to find new ways to maintain production efficiency while ensuring employee safety. Social distancing regulations and labor shortages have prompted many factories to seek automation and remote monitoring solutions.(3) Under these circumstances, traditional factory safety measures face significant challenges, making technological innovation one of the key strategies. According to statistics from Taiwan, the number of occupational deaths and disability incidents exceeded 10,000 cases annually between 2018 and 2022, as shown in Fig. 1.

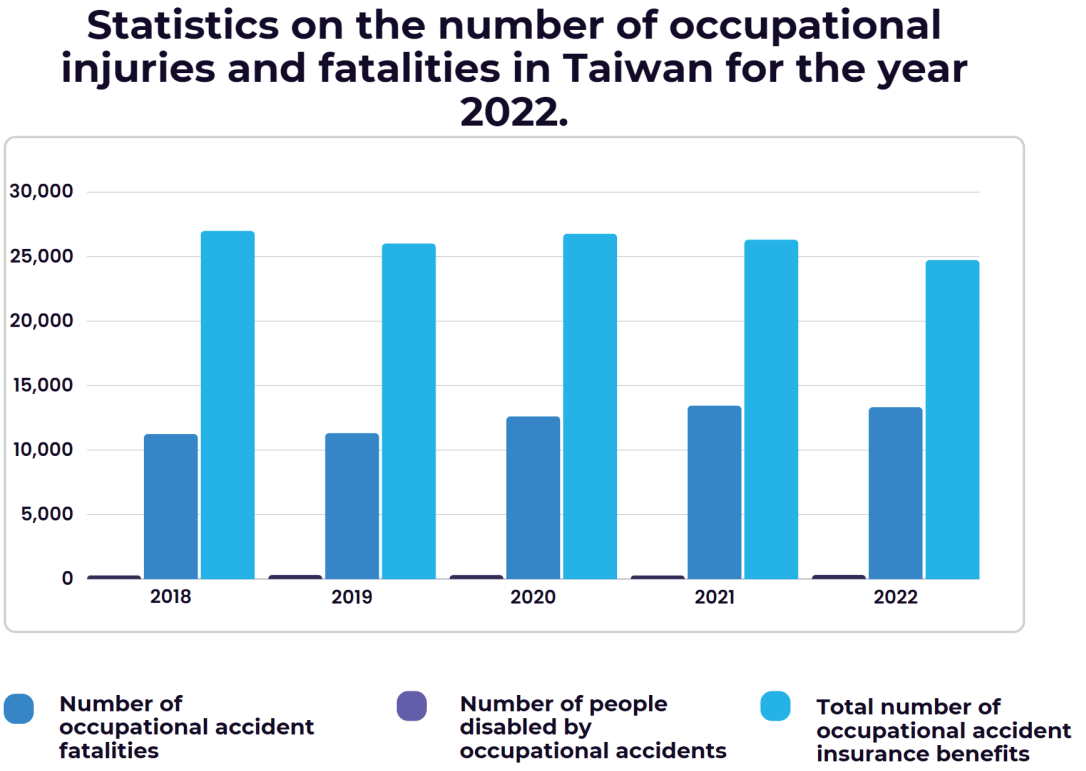

Fig. 1. (Color online) Statistics on the number of occupational injury deaths in 2022 from the Occupational Safety and Health Administration, Ministry of Labor, Taiwan.

In 2022, 320 occupational death cases were recorded. These incidents resulted in significant economic losses to businesses, with direct economic losses amounting to NT$6.68 billion and total economic losses reaching NT$33.4 billion.[4] This highlights the importance of improving occupational safety and advancing technological innovation, especially in an era of increasing automation and intelligent manufacturing. In fact, many industrial accidents are related to improper or incorrect human operations. These accidents can lead to significant economic losses and even threaten workers' lives. However, the automation of industrial production lines and equipment has become an irreversible trend.[5] When problems occur in industrial machine operations, they can not only cause harm to operators but also cause significant damage to equipment and the production environment. According to statistics from the Occupational Safety and Health Administration of Taiwan's Ministry of Labor, accidents caused by unsafe actions combined with equipment account for 44% of total disability injuries, accidents solely caused by unsafe actions account for 31%, accidents caused by unsafe equipment account for 21%, and other causes account for 4%. These data highlight the importance of enhancing the monitoring and management of machine operations in the trend of automation and intelligent manufacturing, especially in promptly addressing issues to protect operators' safety and prevent further damage to equipment and the production environment.

Therefore, in the monitoring and management of machine operations, we need to consider various scenarios before, during, and after operations to ensure production safety, protect the lives of operators, and quickly resume normal production in case of issues. In a factory safety study, Su *et al.*[6] proposed a method using edge computing and Bluetooth sensor technology to monitor changes in the factory environment. This system can trigger alarms and automatically control the production line's programmable logic controllers to perform corresponding actions when environmental data exceed preset thresholds, thereby enhancing real-time safety monitoring capabilities and addressing potential risks.

To address this challenge, in this study, we focus on advanced computer vision technologies, selecting cutting and press machines as research cases. By examining machines prone to accidents in the industry, we explore how to effectively apply artificial intelligence, machine learning techniques [particularly convolutional neural networks (CNNs)], and sensor technologies. This approach not only improves operational accuracy but also enables the real-time identification and prevention of potential safety risks. Additionally, with the application of IoT technology, these safety monitoring systems can achieve remote monitoring and data management, providing real-time feedback and decision support for factory management, thereby ensuring safety and sustainability while maintaining production efficiency.

The combination of AI and IoT technologies is a key factor driving industrial automation and intelligent manufacturing.[7,8] In particular, deep learning, as a branch of machine learning, provides a powerful method for analyzing and interpreting complex data, especially in image recognition and behavior analysis. CNN, as a type of deep learning, can learn and identify unsafe behavior patterns from images captured by cameras and provide immediate feedback. However, achieving such a high level of automation and real-time monitoring requires more than just AI. This is where IoT technology plays a crucial role. IoT refers to a network of interconnected physical devices that can extend AI's capabilities, enabling real-time data collection and analysis on a broader and deeper scale, as shown in Fig. 2.

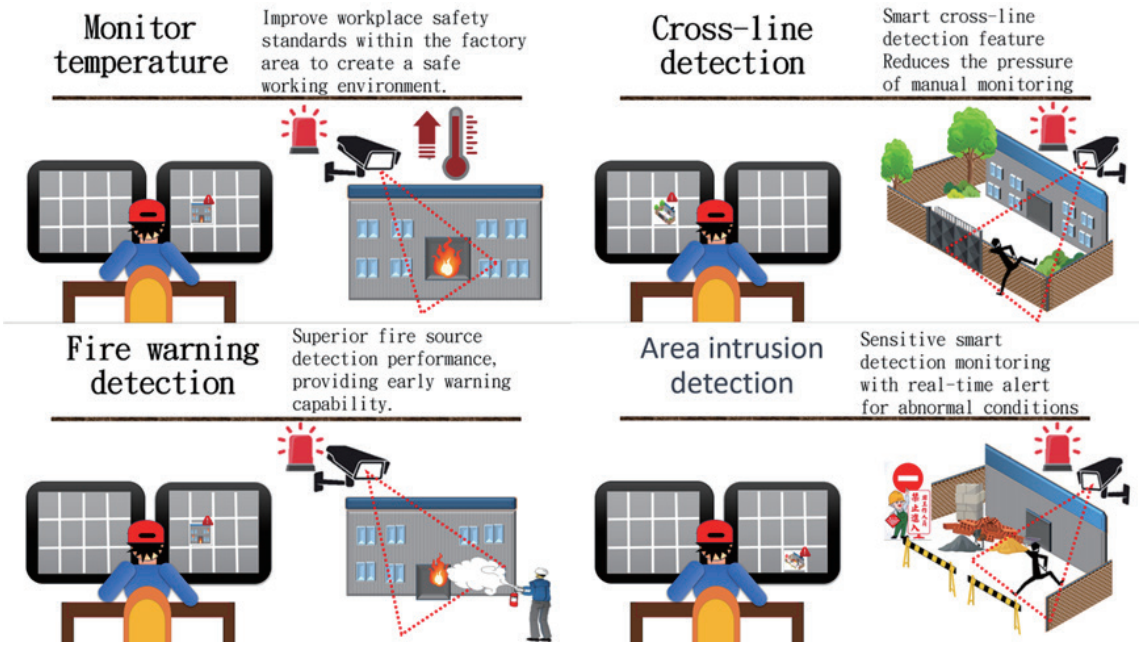

Fig. 2. (Color online) Extensive application of IoT in industry.

In factory safety monitoring systems, IoT can connect cameras, sensors, and other devices, transmitting the collected data to AI models for analysis. This integration not only enhances the system's monitoring scope and precision but also improves the speed and accuracy of responses to potential hazards.

The goal of this research is to develop and validate an innovative factory safety monitoring system based on an improved CNN integrated with IoT technology. We specifically focus on cutting and press machines owing to their widespread use and higher potential risk in industrial manufacturing processes. By conducting an in-depth study of the operational processes and potential hazardous behaviors associated with these machines, we aim to develop an intelligent system capable of identifying and preventing such hazards. This research will focus on three core areas: first, improving and optimizing existing CNN models by introducing 3D image recognition to precisely identify and predict unsafe behaviors in the factory, ensuring the system's capability to efficiently process and analyze video data; second, integrating IoT technology for data collection, transmission, and processing, further expanding the scope and efficiency of the safety monitoring system; and finally, deploying and evaluating the system in real factory environments to validate the model's effectiveness in the real-time monitoring and prevention of industrial accidents.

In Sect. 2, we will discuss the application of 3D CNN in factory safety monitoring. In Sect. 3, we will explain the sensor technology and data collection methods used in this research. In Sect. 4, we will discuss the implementation of IoT and factory safety monitoring. Finally, in Sect. 5, we will provide conclusions to this study.

2. Application of 3D CNN in Factory Safety Monitoring

In today's industrial environment, factory safety monitoring is a crucial issue. Traditional monitoring methods often rely on manual operations and basic automation technologies, which may not effectively identify and prevent various industrial safety risks. With the development of machine learning techniques, particularly deep learning technologies such as CNNs, there has been widespread interest in the field of industrial safety monitoring. As an important branch of machine learning, CNNs can automatically learn and extract features, significantly improving the accuracy and efficiency of identification.[9,10] Specifically, 3D CNNs can not only learn spatial features but also capture dynamic features that change over time,[11] making them particularly effective in processing video and dynamic scenes. This is critical for monitoring and preventing safety incidents in factories. For example, 3D CNNs can effectively identify abnormal behaviors in videos, such as falls or dangerous operations, and provide timely alerts.[12]

A CNN is composed of numerous neurons, each of which has inputs and outputs. The inputs come from the neurons in the previous layer, while the outputs are passed on to the neurons in the next layer. The connections between neurons are represented by weights, which indicate the strength of signal transmission. The neural network processes data by inputting it into the neurons, performing a series of calculations, and adjusting the weights to maximize output accuracy. This involves a convolutional kernel, known as a local receptive field, which slides over the units of the previous layer. Each unit within the convolutional layer receives input from a set of units within the convolution kernel window and calculates on the basis of the following equation:

$$f_{xy} = \tanh\left(\sum_{i,j} w_{ij} v_{(x+i)(y+j)} + b \right), \tag{1}$$

where f_{xy} represents the feature map unit at position (x, y), $v_{(x+i)(y+j)}$ is the input unit at position $(x + i)(y + j)$, $\tanh(\cdot)$ is the activation function, b denotes the bias, and w_{ij} are the weights of the convolution kernel.

Neural networks have various structures, such as feedforward neural networks (FNNs), recurrent neural networks (RNNs), and CNNs.[13–15] Each structure has its own characteristics and applicable scenarios. FNNs are the simplest form of artificial neural network architecture, with no cyclic connections between nodes, making them suitable for a wide range of tasks from simple to complex functions. RNNs are characterized by their internal memory, allowing them to process input sequences, making them well-suited for tasks involving sequential data, such as time series analysis, natural language processing, and speech recognition. RNNs can handle the dynamic changes in time series data, learning patterns that evolve over time.[16]

2.1 CNNs

Among these structures, CNNs are specifically designed for processing data with a grid-like topology, such as images. CNNs typically consist of four types of layer: convolutional, pooling,

activation, and fully connected layers. A distinctive feature of CNNs is their convolutional layers, which extract features from input data and can automatically and adaptively learn spatial feature hierarchies from input images.[17,18] This characteristic is particularly suited for image recognition tasks, allowing the network to become increasingly accurate as training progresses. This enables CNNs to automatically identify key elements in the input data without requiring manual intervention. By configuring filters to detect specific patterns, CNNs become a deep learning model specifically designed for processing and analyzing 3D data. In the convolutional layers used for learning 3D data, 3D convolutions are performed to capture features in both spatial and temporal dimensions. This method has shown excellent performance in real-time object recognition and classification, significantly improving accuracy and processing speed in object recognition applications.[19] During computation, consecutive frames are stacked into a cube and convolved with a 3D kernel. The computation of 3D convolution is as follows:

$$f_{xyt} = \tanh\left(\sum_{i,j,k} w_{ijk} \upsilon_{(x+i)(y+j)(t+k)} + b\right). \tag{2}$$

In this way, the feature maps of the convolutional layer connect the consecutive frames from the previous layer, capturing motion information. Each 3D convolutional kernel extracts a single type of feature from the frame cube, as the kernel weights are replicated throughout the cube. According to the design principles of CNNs, the number of feature maps in subsequent layers is increased by generating multiple features from the same low-level feature maps. This can be achieved by applying multiple 3D convolutions (using different kernels) at the same location of the previous layer, similar to the approach in 2D convolutions.

Compared with traditional 2D CNNs, 3D CNNs have the main advantage of being able to automatically learn and extract features without the need for manual feature engineering.[20] The difference between the two is illustrated in Fig. 3.

In image recognition, CNNs can detect specific patterns, such as vertical or horizontal lines, through specially configured filters. When set to a vertical line filter, CNNs can identify and extract all vertical lines in an image; similarly, when set to a horizontal line filter, it focuses on identifying all horizontal lines. These filters are applied across the entire image through convolution operations, effectively extracting linear features from the image. This method allows CNNs to precisely identify and classify specific graphical features in complex images, supporting more advanced image recognition tasks. When processing data with temporal sequence characteristics, such as videos or 3D medical images, 3D CNNs can capture richer spatiotemporal information.[21]

3D CNNs extend the functionality of traditional CNNs by performing convolution operations along the time dimension, learning not only spatial features but also dynamic features that change over time, adding a depth consideration. This allows them to capture spatial features that evolve over time. In terms of filter application, 3D CNN filters move and convolve not only in the horizontal and vertical directions of the image but also along the time axis. This means that 3D CNNs can identify and analyze structures and movements that change over time, such as identifying motion patterns or changes between consecutive frames in a video. This capability is used to process dynamic images and understand complex temporal sequence data, making 3D

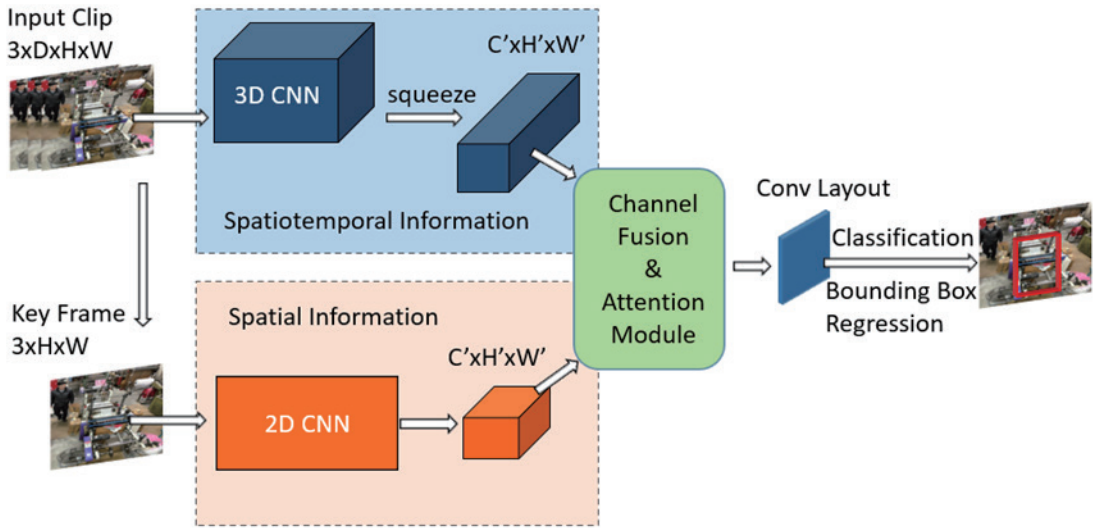

Fig. 3. (Color online) Comparison between 2D and 3D CNNs.

CNNs particularly effective in identifying and understanding continuous actions and behaviors in videos. With the application of visualization techniques, there have been innovations in feature extraction and architecture that differ from traditional CNNs.[22]

Szegedy *et al.* proposed the VGG based on a simple topological structure and introduced GoogLeNet (Inception), which allows the identification of patterns of different sizes within the same layer.[23] Xie *et al.* significantly improved performance by introducing impactful residual blocks that combine the outputs of skip connections and convolutional blocks through addition.[24] Further optimizations to architectures such as Inception-v4 and ResNeXt by Huang *et al.* have further enhanced the efficiency and accuracy of neural networks.[25]

DenseNets and Dual Path Networks (DPNs) demonstrate the powerful capabilities of modularity and structural combination. To accommodate real-time applications and constrained hardware,[26] MobileNets introduced depthwise separable convolutions, providing a more flexible and efficient solution.[27] This deep learning technology offers a powerful tool for applications in factory safety monitoring, providing strong capabilities for detecting and preventing dangerous behaviors by operators through its ability to process temporal sequence features.

Effective research and development require not only theoretical support but also practical cases to verify the feasibility and effectiveness of the technology. In recent years, the application of deep learning technology in industrial safety monitoring has gradually gained attention. For example, one study proposed a model capable of detecting abnormal worker behaviors in manufacturing environments, thereby ensuring the safety of industrial production.[28] Another study explored how to use deep learning technology to prevent industrial accidents, particularly focusing on safety issues when personnel operate machinery.[29] Zhao *et al.*, in their paper, emphasized human behavior recognition in videos.[30] They explored how to address the poor performance of CNNs in capturing temporal features by using a traditional two-stream CNN

network. The researchers attempted to design and implement a new two-stream model that used an LSTM-based model in its spatial stream to extract spatiotemporal features from RGB frames. However, for certain specific behavior recognition tasks, the two-stream model might not be as effective as other single models.[30] Song *et al.* proposed a Modality Compensation Network, aiming to explore the relationships between different modalities and enhance the representation of human behavior recognition.[31] The goal of this study was to extract more discriminative features from the source modality, with RGB/optical flow videos as the source modality and skeletons as the auxiliary modality, only requiring the assistance of skeletons during the training phase. However, in practical applications, there might be issues of modality mismatch that need to be addressed.[31]

The most important deep learning architectures have been validated through the ImageNet Large-dcale Visual Recognition Challenge. One of the earliest winners of this competition was AlexNet, followed by ZFNet and OverFeat, which improved their results on the ImageNet dataset.[20,32,33]

2.2 Transfer learning in 3D CNN

Transfer learning refers to the process of fine-tuning or specializing the feature representations learned by a neural network pretrained on a specific dataset to adapt to a new target dataset. Research in video classification and understanding has been ongoing for decades, starting from handcrafted features such as HOG3D and iDTs to the introduction of deep learning,[34,35] which has made the learning of spatiotemporal features from 2D to 3D data automatic and efficient. These methods are often used to leverage existing extensive learning knowledge to improve the learning efficiency and performance of new tasks. 3D CNNs are particularly outstanding in processing videos and dynamic scenes, capturing dynamic features of time series through 3D convolution.[36] Transfer learning further enhances the performance of these models by allowing pretrained 2D networks to adapt to 3D data. Recent studies have shown that transfer learning can effectively improve model performance within the same modality and across modalities (e.g., from RGB images to RGB images, from RGB images to depth information, from RGB images to optical flow, from RGB images to sound, and from near-infrared images to RGB images).[37]

A recent study[38] proposed a model called RD3D, which applied 3D CNN to RGB-D saliency object detection for the first time. The RD3D model performs prefusion in the encoder stage and deep fusion in the decoder stage, effectively improving the integration of RGB and depth data, and significantly enhancing detection accuracy. Additionally, another study[39] explored how to transfer the weights of pretrained 2D Convolutional Neural Networks (2D CNNs) to 3D CNNs to improve the accuracy of industrial part classification. These techniques demonstrate the powerful capability of 3D convolution in processing multimodal data and applying transfer learning, further proving the broad potential of 3D CNNs in industrial and monitoring applications, and emphasizing the potential of transfer learning in enhancing the performance of deep learning models across different domains.

Some researchers, leveraging the fact that 2D deep learning is more mature than 3D deep learning, have attempted to seek 3D solutions based on 2D deep learning. Recently, with the advent of RGB-D sensors such as Microsoft's Kinect or Intel's RealSense, obtaining 3D information at a low cost has become possible. These sensors provide 2D color images (RGB) and depth maps (D), the latter of which provides 3D information. Since both RGB and D are 2D images, 2D deep learning methods can be adapted to receive two images as input instead of one. Although this representation is relatively simple, it is effective for different tasks, such as human pose regression, 6D pose estimation, and object detection.[40–42]

2.3 Comparison of 3D CNNs with other common neural networks

In the field of deep learning, CNNs and their extended version, 3D CNNs, are the primary tools for processing image and video data. Compared with other common neural networks, such as FNNs and RNNs, CNNs and 3D CNNs have significant advantages in handling data with a spatial structure. The key comparisons of these methods in the application of factory safety monitoring are described in Table 1.

CNNs and 3D CNNs: CNNs are primarily used for processing two-dimensional images, capable of automatically learning spatial features within images, and are widely applied in tasks such as image classification and object detection. In contrast, 3D CNNs extend the capabilities of CNNs by handling 3D data, such as videos or RGB-D data, not only learning spatial features but also capturing dynamic features that change over time. This makes 3D CNNs particularly effective in industrial safety monitoring, where they excel in capturing abnormal behaviors (e.g., falls and dangerous operations).

FNNs and RNNs: FNNs are the most basic neural network architecture, suitable for static data and nonsequential problems, but they cannot handle data that changes over time, limiting their effectiveness in dynamic scenarios.[43]

RNNs are capable of processing sequential data, making them well-suited for tasks involving time series analysis, natural language processing, and speech recognition, as they have internal memory that can remember inputs from previous moments. However, RNNs are less effective when dealing with image and video data, particularly when there is a need to process both spatial and temporal features simultaneously.[44]

CNNs and 3D CNNs demonstrate clear advantages in industrial applications, particularly in processing image and video data. FNNs are suitable for static data processing, while RNNs are better at handling time series data. In comparison, 3D CNNs can process both spatial and temporal features simultaneously, making them especially suitable for application scenarios that

Table 1
Comparison of this study with common machine learning methods.

Method	Time complexity	Space complexity	Applicable to static data	Applicable to video data	Anomaly detection capability
FNN	Low	Low	Yes	No	Weak
RNN	High	Medium	No	Yes	Medium
CNN	Medium	Medium	Yes	Limited	Medium
3D CNN	High	High	Yes	Yes	Strong

require multimodal data fusion and real-time monitoring, such as factory safety monitoring.

In this study, we first used 2D images of cutting machine operations as the basis for pretraining, and then transferred and fine-tuned the learned features to adapt to 3D operation videos in a plastic bag factory. Transfer learning allows us to specialize the feature representations learned by pretrained neural networks, thereby improving the learning efficiency and performance of new tasks. This method effectively leverages existing extensive learning knowledge by jointly learning representations of different modalities in a shared feature space. Our goal is to achieve supervised transfer from 2D convolutional networks to 3D convolutional networks, including transfers from RGB images to other modalities, and to use the 2D color images and depth maps provided by RGB-D sensors to obtain 3D information. Although these representations are relatively simple, this 2D-based deep learning approach is effective for different tasks and can be flexibly adapted for multimodal learning. Through these strategies, 3D CNNs excel in processing videos and dynamic scenes, effectively capturing dynamic features of time series and significantly enhancing model performance in factory safety monitoring.

3. Sensor Technology and Data Collection

In this study, we used high-definition (720p, 30 frames per second) Xiaomi Cloud Cameras connected via WiFi to a server equipped with a 32-core Intel x86 processor, an advanced RTX 4090 graphics card, and the TensorFlow framework. This setup demonstrates the immense potential of integrating IoT with advanced imaging sensor technology, especially in enhancing industrial safety and efficiency. This configuration not only allows us to monitor specific environments in real time but also automatically extracts key information from images and videos through computer vision technologies, such as 3D models, camera positioning, object detection, and recognition. This technology integrates expertise from fields such as image processing, pattern recognition, mathematics, and artificial intelligence, enabling the system to perform core visual functions such as motion sensing and scene understanding.[45] Importantly, this hardware is relatively accessible for businesses, including general users, and the cost of deploying this technology throughout a factory environment is not overly burdensome for most businesses, with relatively easy maintenance.

We employed an improved version of the ViBe algorithm, a classic background subtraction algorithm used for detecting and segmenting moving objects. This algorithm is primarily applied to detect various vehicle types, crowds, pedestrians, and personnel access control, compensating for the limitations of operators to monitor numerous cameras simultaneously. Recent research, such as that by Gan and Zhao, has made improvements to the ViBe algorithm.[46] By enhancing the ViBe algorithm and integrating various feature fusion methods, they effectively removed shadows from images and improved the accuracy of moving object detection, thus enhancing performance in complex backgrounds. In another study, Yu *et al.*[47] improved the initialization process of the ViBe algorithm, reducing ghosting effects caused by moving objects in the first frame. These enhancements have made the ViBe algorithm a more accurate tool for identifying and tracking both persistent and transient objects in specific environments within intelligent

visual monitoring systems. Combined with powerful computational resources and IoT technology, our system is capable of capturing rapid movements and subtle changes, as well as transmitting data in real time to servers for analysis and model training.

Furthermore, since the resurgence of neural networks, deep learning models, especially deep neural networks, have made significant progress in digital image processing areas, such as automatic object and facial recognition. Our system, integrating the image capture capabilities of high-definition cameras with IoT technology, can transmit data in real time to servers for in-depth analysis and model training, thereby effectively improving the accuracy and efficiency of industrial safety monitoring. In plastic bag manufacturing plants, the system is deployed at key locations around cutting and processing machines, specifically monitoring high-risk areas such as sharp blades, moving rollers, and high-temperature heating plates.[48–51]

Overall, our system excels in the real-time monitoring of persistent and transient objects in specific environments. High-definition cameras capture rapid movements and subtle changes, and by combining IoT with powerful computational resources, data is transmitted in real time to servers for analysis and model training, effectively improving the accuracy and efficiency of industrial safety monitoring.

In plastic bag manufacturing plants, the core of the safety monitoring system lies in real-time monitoring and preventing potential industrial accidents. Servers equipped with RTX 4090 graphics cards provide the computational power needed to quickly process vast amounts of video data from multiple cameras, run complex deep learning models, and perform large-scale data analysis. We utilized the TensorFlow framework to train deep learning models that extract useful information from video data captured by cameras, identifying potential safety risks. The application of IoT technology enables smarter and more flexible connections between cameras and servers, allowing for the rapid adjustment of data collection strategies based on the current production status, and integrating data into a centralized platform for comprehensive analysis and processing.

We used the TensorFlow framework to train and deploy specialized deep learning models designed to extract useful information from video data captured by cameras and identify potential safety risks. For example, the models can identify when operators approach hazardous areas, such as the blades of cutting machines or high-temperature heating plates. The application of IoT technology allows for smarter and more flexible connections between cameras and servers. The system can monitor the entire factory environment in real time, collecting data from various production machines such as processing and cutting machines. Through IoT technology, we can quickly adjust data collection strategies on the basis of the current production status and integrate data into a centralized platform for comprehensive analysis and processing.

3.1 Dataset structure

To train deep learning models for factory safety monitoring, we developed a smart monitoring system based on deep learning designed to enhance the efficiency and accuracy of industrial safety monitoring. The system integrates high-definition camera data collection, advanced

image processing technologies, and deep learning models for the real-time identification and prevention of safety risks. Initially, image frames are extracted from videos recorded by cameras, selecting 10000 frames (including 5000 safe and 5000 unsafe images) from a total of 19,200 video segments to provide a balanced and comprehensive dataset, ensuring that the model learns to predict safety risks in various scenarios.

The OpenCV library is used to process video data. A defined "extract_frames" function extracts a fixed number (up to 10 frames) of image frames from each video. Each frame's size is resized to 256 × 256 pixels to ensure data consistency and reduce computational complexity. The "load_dataset" function is then used to load and label the dataset. This function iterates through all video files in a specified path, uses the "extract_frames" function to extract frames from each video, and labels each video frame as "safe" or "unsafe" for subsequent classification training.

The trained model is deployed on the production line to process camera data in real time. It instantly identifies unsafe behaviors during machine operations, such as operators approaching high-risk areas. Additionally, it identifies and responds to subtle safety issues, such as improper operating postures or ignoring safety warnings. Our deep learning model framework includes multiple convolutional, pooling, flattening, and dense layers to effectively process the input 3D image data and make classification predictions. To avoid overfitting, an early stopping strategy is employed during model training, along with data preprocessing and augmentation to improve the model's generalization ability and increase the diversity of training data. Furthermore, we plan to compare the application of 3D CNNs with the 2D CNN method previously proposed for hazard prevention in industrial cutting machines,[52] analyzing their advantages in industrial safety applications.

4. Practical Applications of Factory Safety Monitoring

On a plastic bag automatic processing machine, the main hazardous areas are divided into three parts: the rear, middle, and front sections of the machine, as shown in Fig. 4.

The hazardous parts of the machine include blades for cutting plastic bags, rear rollers for transporting the plastic tape during the feeding process (marked with a black frame), middle sections with high-temperature heating plates for sealing plastic bags (marked with a red frame), and front sections with rollers and cutting blades for discharging and cutting the plastic tape (marked with a yellow frame).

During the operation of plastic bag processing machines, personnel may suffer significant physical injuries if not careful, especially in the three main hazardous areas: the feeding rollers at the rear, the high-temperature heating plates in the middle, and the discharge rollers and cutting blades at the front. The improper operation of the rear rollers may result in hands or clothing being pulled into the rollers, leading to pinching or more severe cutting injuries. Contact with the middle section's high-temperature heating plates can cause burns, as these plates are hot enough to melt plastic. The improper operation of the front cutting blades may lead to cuts or more severe finger amputation accidents.

Fig. 4. (Color online) Aerial view of a roll-to-roll plastic bag processing machine.

The following sequence of video screenshots, captured in just 6 s, demonstrates the transition from routine safe operations to potentially accident-causing dangerous behaviors. This sequence highlights the rapid changes in safety risks within industrial environments and the urgency of preventive measures, as shown in Fig. 5.

This rapid behavioral transition underscores the critical role of 3D CNNs. Unlike traditional 2D CNNs, 3D CNNs can capture essential features of temporal continuity in videos. The sequence of operator actions and interactions with machinery over time in the videos provides crucial clues for predicting potential hazards. 3D CNNs can identify these dynamic features and respond to possible dangerous behaviors in real time. The advantage of 3D CNNs in processing video surveillance data lies in their ability to understand not only the spatial features within a single frame but also the temporal relationships between frames. Safety monitoring requires identifying dangers in static images and, more importantly, recognizing and understanding actions and behavior patterns that change over time.

In our experiment, we compared 3D CNNs with 2D CNNs. We used 3D CNNs to analyze videos recorded from four cameras installed on two production machines. These videos covered

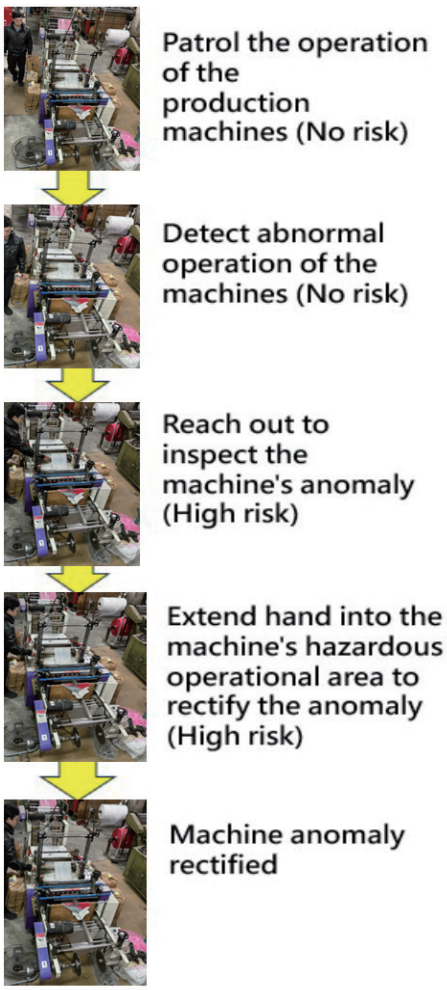

Fig. 5. (Color online) Video sequence showing the evolution from normal operations to potential hazardous behaviors.

8 h of operation and were segmented at a frequency of one segment every 6 s, resulting in a total of 19200 video segments. Since the majority of the time, the machines operate automatically during actual factory operations, and adjustments to the machines or materials are only made when anomalies occur or when defects are found in the products, we carefully selected 4000 frames from the prerecorded video dataset. This selection included 2300 frames depicting normal operations marked as safe and 1700 frames marked as unsafe, specifically chosen for abnormal behaviors that might occur during machine adjustments and actions taken to troubleshoot machine malfunctions, ensuring the dataset's balance and comprehensiveness. This dataset allowed the model to learn and predict safety risks in various scenarios, fully utilizing the advantages of 3D CNNs in learning spatial and temporal sequence features. The proposed model structure is illustrated in Fig. 6, highlighting the integration of spatial and temporal features, which is critical for effective safety monitoring. After training the dataset with our proposed model structure, the results are shown in Fig. 7.

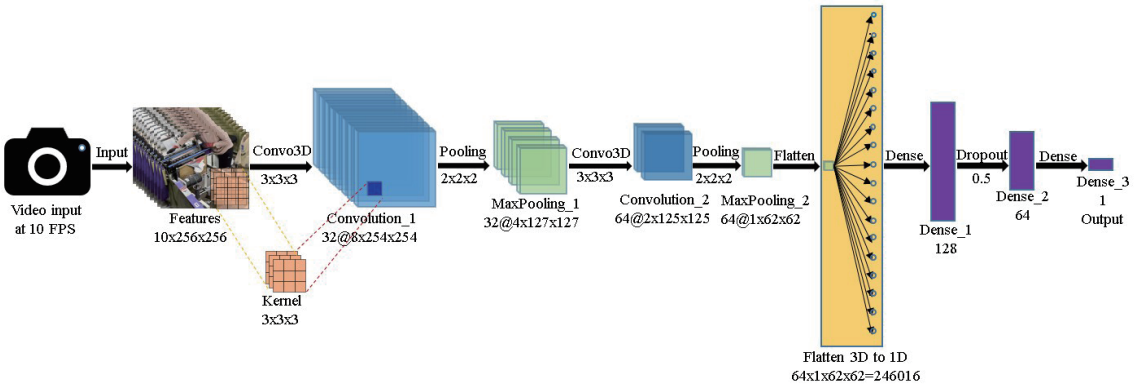

Fig. 6. (Color online) Model structure used in this study.

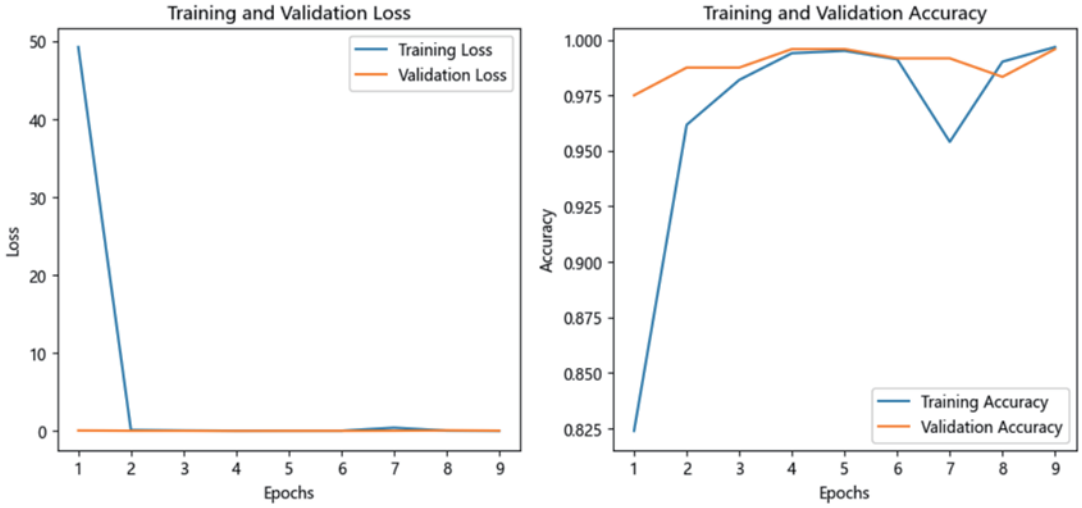

Fig. 7. (Color online) Training and validation losses and accuracies of the 3D CNN in this study.

In the initial epochs, the model rapidly learned, with the training loss significantly decreasing from 49.2363 to 0.0092 by the 9th epoch and the training accuracy improving from 82.39 to 99.67%. This rapid progress indicates that the model effectively learned to distinguish between safe and unsafe operational behaviors. Meanwhile, the validation accuracy steadily increased to 99.58%, demonstrating the model's good generalization ability on unseen data. Although the validation loss began to rise from 0.0197 after the 7th epoch, reaching 0.0362 by the 9th epoch, this upward trend was not significant, and the validation accuracy remained high at 99.58% by the 9th epoch. This reflects the model's high accuracy on validation data, indicating that overfitting was not significant. Furthermore, the use of early stopping mechanisms allowed training to stop at an appropriate time, preventing model overfitting and ensuring both training efficiency and model quality.

For the 2D CNN model, using the same setup of four cameras on two production machines, we selected 8000 static images for training and testing, with 80% used for training and 20% for validation. This data allocation ensured that the model had sufficient data for learning. The experiments showed that it performed excellently on static images, achieving a training accuracy

of 99.93% at 27 epochs, and the best validation accuracy was 97.92%. The validation loss reached its lowest point at 0.0408 during the 22nd epoch, at which point the early stopping mechanism was triggered. Subsequent training epochs failed to further reduce the validation loss, so the model's weights were restored to this optimal state. This confirms the high efficiency of 2D CNNs in classifying static images, particularly in distinguishing between safe and unsafe operations, as illustrated in Fig. 8.

In our experiments, 3D CNNs demonstrated comparable or even superior accuracy to 2D CNNs while also providing a deeper understanding of the temporal dimension, which is crucial for analyzing continuous actions and predicting behavioral patterns. By identifying dynamic features in time series, 3D CNNs can capture subtle changes during operations, which is invaluable for preventing potential safety incidents. Furthermore, by analyzing consecutive frames, 3D CNNs can more comprehensively assess the behavioral risks of operators over time. In real industrial environments, this capability is essential for understanding complex situations that may arise during workflows. Considering all factors, although 2D CNNs have demonstrated excellent performance in static image classification, 3D CNNs offer richer and more in-depth information for analyzing dynamic behaviors and predicting potential safety risks, providing a key technical advantage for early warning and accident prevention in high-risk production environments.

In summary, while 2D CNNs have already shown outstanding performance in static image classification, 3D CNNs provide richer and more comprehensive information for analyzing dynamic behaviors and predicting potential safety risks. This dynamic analysis capability is a key technical advantage for implementing early warnings and preventing accidents in high-risk production environments.

5. Conclusions

The aim of this study was to develop an innovative factory safety monitoring system that integrates 3D CNNs with IoT technology to meet the high demands of modern industrial safety

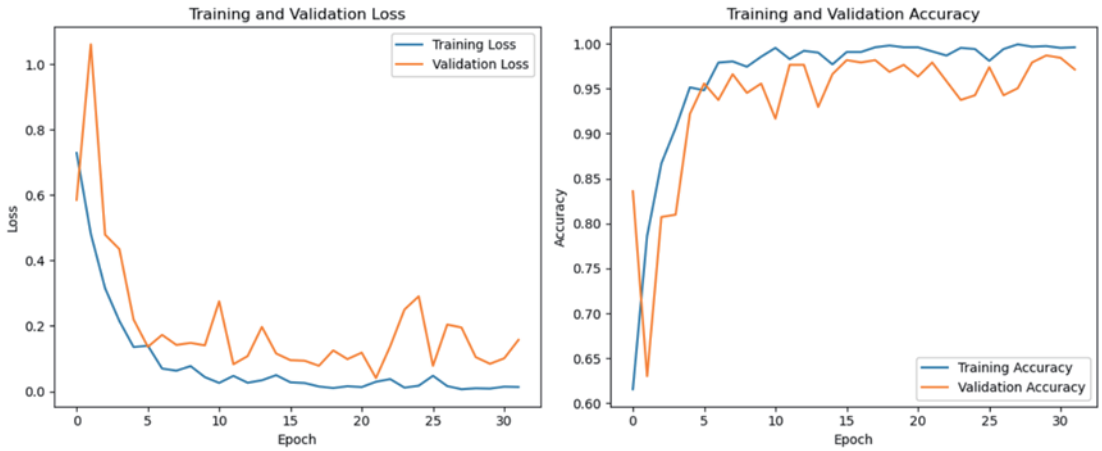

Fig. 8. (Color online) Training and validation losses and accuracies of the 2D CNN in this study.

monitoring. The system's design considered accessibility, ease of deployment, and effectiveness in practical applications. The results indicate that the system performs exceptionally well in the real-time identification and prevention of industrial safety risks. The 3D CNN model, leveraging its capability to learn spatiotemporal features, significantly enhanced the system's ability to provide early warnings in dynamic scenarios. We found that 3D CNNs not only match 2D CNNs in terms of accuracy but also outperform them in certain aspects, especially in understanding continuous actions and predicting behavioral patterns, making them an effective solution for industrial safety monitoring.

The model's training outcomes demonstrated that the 3D CNN can rapidly learn from the data and effectively distinguish between safe and unsafe operational behaviors, proving its excellent learning efficiency. Although there was a slight increase in validation loss during the later stages of training, the application of early stopping mechanisms ensured that the model maintained high accuracy on the validation set and stopped training at the appropriate time. This effectively prevented overfitting and preserved the model's generalization ability.

In contrast, the 2D CNN exhibited outstanding performance in classifying static images, achieving a best validation accuracy of 97.92%. This confirmed the effectiveness of 2D CNNs in static feature extraction and classification. However, in the field of safety monitoring, where understanding and predicting behavior patterns based on time series is essential, 3D CNNs provide a more comprehensive solution. By capturing continuous actions in videos, 3D CNNs can offer timely responses to potential hazardous behaviors, a capability not achievable by 2D CNNs.

While 2D CNNs have shown excellent performance in static image classification, 3D CNNs offer richer and more in-depth information for analyzing dynamic behaviors and predicting potential safety risks. This demonstrates their strong potential in applications beyond static image classification. Future work will continue to explore the application of 3D CNNs in various industrial settings and investigate how to further optimize the model to adapt to diverse and complex scenarios. Our system is expected to become a key technology for early warning and accident prevention in industrial environments, offering robust support for industrial safety management.

Acknowledgments

This work was supported by the Ministry of Education, Taiwan, under the Industry College Program, as approved by Official Letter No. 1132315951U, Tai-Edu-Tech (3). The authors are thankful for the funding and resources provided, which significantly contributed to the completion of this research.

References

1 I. Madabhavi, M. Sarkar, and N. Kadakol: Monaldi Archives for Chest Dis. **90** (2020).
2 P. Arunprasad, C. Dey, F. Jebli, A. Manimuthu, and Z. El Hathat: Benchmarking: An Int. J. **29** (2022) 3333.
3 R. Diab-Bahman and A. Al-Enzi: Int. J. Sociol. Soc. Policy **40** (2020) 909.
4 Ministry of Labor, Taiwan: Analysis of industrial damage in Taiwan from 2018 to 2022. https://www.osha.gov.tw/48110/48331/48333/48339/150732/ (Accessed Apr. 2024).

5 A. Chernoff and C. Warman: Appl. Econ. **55** (2023) 17.
6 L. Su, Y. H. Lee, Y. L. Chen, H. W. Tseng, and C. F. Yang: Sens. Mater. **35** (2023) 1731.
7 P. Radanliev, D. De Roure, R. Nicolescu, M. Huth, and O. Santos: Int. J. Intell. Robot. Appl. **6** (2022) 171.
8 S. K. Panda, R. K. Mohapatra, S. Panda, and S. Balamurugan, Eds.: The new advanced society: Artificial Intelligence and Industrial Internet of Things Paradigm (John Wiley & Sons, 2022).
9 C. I. Nwakanma, F. B. Islam, M. P. Maharani, J. M. Lee, and D. S. Kim: Appl. Sci. **11** (2021) 3662.
10 S. Hong, T. Feng, J. Hu, and X. Zhang: IEEE Systems, Man, and Cybernetics Magazine **9** (2023) 4.
11 S. Urabe, K. Inoue, and M. Yoshioka: Proc. Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management (2018) 1–8.
12 J. Su, P. Her, E. Clemens, E. Yaz, S. Schneider, and H. Medeiros: Proc. 18th IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS, 2022) 1–8.
13 S. Albawi, T. A. Mohammed, and S. Al-Zawi: Proc. 2017 Int. Conf. Eng. Technol. (ICET, 2017) 1–6.
14 G. E. Hinton, S. Osindero, and Y. W. The: Neural Comput. **18** (2006) 1527. https://doi.org/10.1162/neco.2006.18.7.1527
15 L. R. Medsker and L. Jain: Design Applic. **5** (2001) 64.
16 C. J. Spoerer, P. McClure, and N. Kriegeskorte: Front. Psychol. **8** (2017) 1551.
17 M. M. Taye: Computation **11** (2023) 52.
18 D. Maturana and S. Scherer: 2015 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS, 2015) 922.
19 T. Epelbaum: arXiv preprint (2017) arXiv:1709.01412.
20 G. Rouhafzay, A. M. Cretu, and P. Payeur: Sensors **21** (2020) 113.
21 T. Nguyen, B. S. Hua, and N. Le: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th Int. Conf. (France, 2021) Proc. Part I, **24** (2021) 548–558.
22 K. He, X. Zhang, S. Ren, and J. Sun: Proc. IEEE Conf. Computer Vision and Pattern Recognition (USA, 2016) 770–778.
23 C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi: Proc. AAAI Conf. Artificial Intelligence **31** (USA, 2017) 4–9.
24 S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He: Proc. IEEE Conf. Computer Vision and Pattern Recognition (USA, 2017) 1492–1500.
25 G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger: Proc. IEEE Conf. Computer Vision and Pattern Recognition (USA, 2017) 4700–4708.
26 Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng: arXiv preprint (2017) arXiv:1707.01629.
27 A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam: arXiv preprint (2017) arXiv:1704.04861.
28 F. Franzel, T. Eiband, and D. Lee: 2020 IEEE-RAS 20th Int. Conf. Humanoid Robots (Humanoids, 2021) 376.
29 A. Bejaoui, C. He, and D. Soffker: Annu. Conf. PHM Society **14** (2022).
30 Y. Zhao, K. L. Man, J. Smith, K. Siddique, and S. U. Guan: EURASIP J. Image and Video Proc. (2020) 1–9.
31 S. Song, J. Liu, Y. Li, and Z. Guo: IEEE Trans. Image Process. **29** (2020) 3957.
32 A. Krizhevsky, I. Sutskever, and G. Hinton: Adv. Neural Inf. Process. Syst. **25** (2012) 1097.
33 Zeiler and R. Fergus: European Conf. Computer Vision (Switzerland, 2014) 81.
34 A. Klaser, M. Marszałek, and C. Schmid: 19th British Machine Vision Conf. (British Machine Vision Association, 2008) pp. 275:1–10.
35 H. Wang and C. Schmid: Proc. IEEE Int. Conf. Computer Vision (IEEE, 2013). 3551–3558.
36 S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy: Proc. European Conf. Computer Vision (ECCV, 2018) 305–321.
37 A .Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool: arXiv preprint (2017). arXiv:1711.08200.
38 Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du: Proc. AAAI Conf. Artificial Intelligence **35** (2021) 1063.
39 I. Merino, J. Azpiazu, A. Remazeilles, and B. Sierra: Sensors **21** (2021) 1078.
40 G. Fanelli, T. Weise, J. Gall, and L. Van Gool: Joint Pattern Recognition Symp. (Germany, 2011) 101.
41 J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birch-field: arXiv preprint (2018) arXiv:1809.10790.
42 N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu: Proc. IEEE Conf. Computer Vision and Pattern Recognition (USA, 2015) 362–370.
43 O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad: Heliyon **4** (2018) e00938.
44 Y. Yu, X. Si, X. C. Hu, and J. Zhang: Neural comput. **31** (2019) 1235.
45 B. Pang, E. Nijkamp, and Y. N. Wu: J. Educ. Behav. Stat. **45** (2020) 227.

46 L. Gan and H. X. Zhao: Microelectron. Comput. **32** (2015) 152.
47 Y. Yu, H. Pan, X. Li, and J. Lin: 2018 IEEE 8th Annu. Int. Conf. CYBER Technology in Automation, Control, and Intelligent Systems (CYBER, 2018) 445.
48 M. He, B. Li, and H. Chen: 2017 IEEE Int. Conf. Image Processing (ICIP, 2017) 3904.
49 T. S. Borkar and L. J. Karam: IEEE Trans. Image Processing **28** (2019) 6022.
50 S. Maitra, U. Bhattacharya, and S. K. Parui: 2015 13th Int. Conf. Document Analysis and Recognition (ICDAR, 2015) 1021–1025.
51 M. MaYing and K. C. Kim: J. Korea Inst. Electron. Commun. Sci. **14** (2019) 1207.
52 C. L. Chiang, M. Y. Peng, I. L. Lin, Y. W. Chou, J. M. Fong, and Y. Y. Chiang: 2023 IEEE 3rd Int. Conf. Electronic Communications, Internet of Things and Big Data (ICEIB, 2023) 487.

About the Authors

Chien-Lin Chiang is currently a doctoral student at Tatung University. Since 2022, he has served as a lecturer at Vanung University. He is also the secretary of the Taiwan Association for Digital Forensics Development (ACFD) and the Vice General Manager of Baofeng Plastics Company. His research areas include artificial intelligence, computer hardware integration, blockchain applications, and information security. He is dedicated to promoting automation and transformation in traditional industries, as well as industrial safety applications in his field. (f2267505@gmail.com)

I-Long Lin received his B.S. degree from Central Police University, Taiwan, in 1983 and his M.S. and Ph.D. degrees from Tamkang University and National Taiwan University of Science and Technology, Taiwan, in 1989 and 1998, respectively. From 1983 to 2011, he worked as a professor at Central Police University, Taiwan. From 2012 to 2021, he was a professor at Yuanpei University of Medical Technology, Taiwan. Since 2021, he has been a professor at Tatung University. His research interests include digital evidence and forensics, and cybersecurity. (cyberpaul@gm.ttu.edu.tw)

Ming-Yuan Peng received his Bachelor of Laws degree from Fu Jen Catholic University and his Master of Laws degree in Maritime Law from National Taiwan Ocean University. At present, he is a Ph.D. candidate at the Graduate Institute of National Development, Chinese Culture University. He is currently serving as a full-time lecturer at Taipei University of Marine Technology. His past positions include Director of the Office for the Top Ten Outstanding Young Persons Foundation and Standing Director of the Chinese HACCP Food Control System Association. He has published seven specialized books, authored 28 journal articles, and holds six patents. He is currently engaging in interdisciplinary learning to develop a secondary expertise in Information Security Management, participating in several Ministry of Education's industry–academia collaboration platforms for information studies, and has obtained the Environmental, Social, and Governance Sustainability Report certification. (f0931@mail.tumt.edu.tw)

Wen-Hsin Liang received her B.A. degree from Shih Chien University, Taiwan, in 2000. Since 2023, she has been a graduate student of the Department of Computer Science and Information Engineering of Vanung University. (thea.liang@gmail.com)

Yi-Yuan Chiang is an assistant professor at Vanung University, Taiwan, where he has been a part of the faculty since 2008. He received his Ph.D. degree in 2007 from Yuan Ze University, Taiwan. His research focuses on machine learning, sensor fusion, and robot systems integration. Notably, he has made significant contributions to the field and holds numerous patents for his innovative work. (yychiang@mail.vnu.edu.tw)