The following publication Liu, P., Chi, H. L., Li, X., & Guo, J. (2021). Effects of dataset characteristics on the performance of fatigue detection for crane operators using hybrid deep neural networks. Automation in Construction, 132, 103901 is available at https://doi.org/10.1016/j.autcon.2021.103901.

Effects of Dataset Characteristics on the Performance of Fatigue Detection for Crane Operators Using Hybrid Deep Neural Networks

Abstract

Fatigue of operators due to intensive workloads and long working time is a significant constraint that leads to inefficient crane operations and increased risk of safety issues. It can be potentially prevented through early warnings of fatigue for further appropriate work shift arrangements. Many deep neural networks have recently been developed for the fatigue detection of vehicle drivers through training and processing the facial image or video data from the public driver's datasets. However, these datasets are difficult to directly use for the fatigue detections under crane operation scenarios due to the variations of facial features and head movement patterns between crane operators and vehicle drivers. Furthermore, there is no representative and public dataset with the facial information of crane operators under construction scenarios. Therefore, this study aims to explore and analyse the features of multi-sources datasets and the corresponding data acquisition methods which are suitable for crane operators' fatigue detection, further providing collection guidelines of crane operators dataset. Variations on public datasets such as real or pretend facial expression, the segment level of human-verified labelling, camera positions, acquisition scenarios, and illumination conditions are analysed. A hybrid learning architecture is proposed by combining convolutional neural networks (CNN) and long short-term memory (LSTM) for fatigue detection. In order to establish a unified evaluation criterion, the effort of the study includes relabelling three public vehicle drivers datasets, NTHU-DDD, UTA-RLDD, and YawnDD, with human-verified labels at the frame and minute segment levels, and training the corresponding hybrid fatigue detection models accordingly. The average detection accuracies and losses are identified for the trained

models of UTA-RLDD, NTHU-DDD, and YawnDD individually. The trained models are used to evaluate the fatigue status of facial videos from licensed crane operators under simulated crane operation scenarios. The results suggest the necessary considerations of different influential factors for establishing a large and public fatigue dataset for crane operators.

Keywords: Tower Crane Operator, Construction Safety, Fatigue Detection, Multi-Sources Datasets, Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM)

1. Introduction

The fatigue of crane operators is one of the key reasons to cause unsafe operations, further leads to construction accidents. In construction operations, tower cranes are essential hoisting resources for enabling the mobility of the project [1]. With the development of prefabricated buildings, prefabricated products have become more and more complicated [2]. These products evolve from light-weight components or large and heavy modules to more substantial and more cumbersome pre-acceptance integrated units [2]. Given this course of prefabricated product evolution, cranes perform a decisive role in the assembly of prefabricated products by lifting them vertically and horizontally [3]. Although cranes are crucial components in many construction operations, they are also accompanied by a significant fraction of construction deaths. Estimates suggest that cranes are involved in up to one-third of all construction and maintenance fatalities [4]. Furthermore, operators' unsafe behaviour is the main reason leading to crane safety issues, especially inadequate training and fatigue of practitioners, causing unsafe practices of crane operations [5]. About 60.5% of the crane operators will continue to work even feeling fatigued due to long working hours, lack of rest breaks, and demanding physical works. 52.6% of the crane operators have not been arranged breaks every working day [5]. The accurate operations and judgments of the crane operator are compromised in securing safety and productivity, particularly in the construction site, due to the high level of congestion and dynamics [6]. Therefore, it is potentially beneficial to automatically detect and warn the fatigue or drowsiness of crane operators, which can support crane operators, site superintendents and safety directors to make the proper shifts and breaks arrangement.

Although fatigue or drowsiness detection is an important research topic and successful solutions have been applied in domains such as vehicle driving, few studies have developed fatigue detection and warning systems for crane operators. Nowadays, there are several groups of techniques for fatigue detection [7-10]: scale measurements, performance measurements, physiological measurements, and behavioural measurements. As for scale measurements, the level of fatigue is evaluated based on the driver's self-estimation, and a subjective estimate termed Karolinska Sleepiness Scale (KSS) [11] is introduced. The KSS relies on answers provided by the subject at a particular time, and it will not accurately and timely measure the slight changes of fatigue [12].

In the crane operations, performance measurements and physiological measurements include trolley movement and jib rotation speed, loads path deviation, heart rate, electrooculogram (EOG), electroencephalogram (EEG), electromyogram (EMG), electrocardiogram (ECG) and so on [6]. Although physiological measurements are highly correlated with the operator's mental state and are most sensitive to fatigue detection, they require operators to wear necessary sensory devices, which is an extra burden and inconvenience for operators. As for performance measurements, they are greatly affected by external factors and the operator's operation habits. Therefore, these methods are intrusive and might not be easy to implement under crane operation scenarios.

Behavioural measurements are obtained from facial movements and expressions using non-intrusive sensors like cameras. The fatigue detection based on computer vision technologies to recognize facial expressions like eye blinks, yawning, and nodding has high accuracy and no impact on the operators' work. This kind of approach can be used to analyse the facial features extracted from the facial videos/images. It performs a high accuracy after the boosting of the development of various deep neural networks. These deep learning approaches facilitate the computer to learn by itself for capturing the key features.

Previous works on fatigue or drowsiness detection focus on extreme fatigue with apparent signs such as yawning, nodding off, and prolonged eye closure [13]. For example, Zhang et al. [14] adopted the convolutional neural network (CNN) to detect yawning by using the features in the nose region instead of the mouth area due to the head turnings of vehicle drivers. However, for crane operators, such explicit signs may not appear until only moments before the accident. Therefore, it is necessary to detect fatigue early to provide more time for crane operators, site superintendents, and safety directors to make proper reactions. On the other hand, previous works on fatigue detection produced results on datasets that include pretending or acted fatigue, like NTHU-DDD in a simulated driving environment [15] and YawDD in a real driving environment [16], or real fatigue, like UTA-RLDD in actual daily life [13]. The "acted" fatigue means the facial images were captured when subjects were instructed to simulate fatigue or drowsiness, compared to "real" fatigue. Besides, different datasets have various collection methods, testing environments, and label modes. It is difficult to compare the accuracy in fatigue detection among the multi-sources' datasets. Furthermore, there is no large, public, and realistic dataset on crane operators. The primary challenge is to determine which kinds of available dataset's characteristics and the collection methods are most suitable for crane operators fatigue detection.

This study develops a hybrid learning architecture to explore and analyze which kind of available dataset's characteristics and the corresponding data acquisition method are suitable for crane operators' fatigue detection. It is designed by combining CNN and long short-term memory (LSTM) for fatigue detection. Firstly, this hybrid learning architecture is adopted for training on three available datasets relabelled by the authors: NTHU-DDD, UTA-RLDD, and YawDD. Then the trained models are used to evaluate licensed crane operators' facial videos captured during simulated crane operations.

The objectives of this study are: (1) to compare the fatigue detection performance on different datasets with real or pretend facial expression; (2) to compare the fatigue detection performance on different human-verified labels at the frame and minute segment levels; (3) compare the fatigue detection performance on different face poses or camera install positions; (4) to explore and analyse which kind of dataset characteristics and the corresponding data acquisition scenario are suitable for crane operators' fatigue detection; (5) to give guidance for building up a large and public realistic fatigue dataset for crane operators, particularly in tower crane operations.

This paper is organized into the following sections: Section 2 demonstrates the results of the literature review. The multi-sources datasets with different scenarios for fatigue detection are then presented in Section 3. Section 4 proposed the framework of hybrid deep neural networks for fatigue detection. Section 5 explains the detailed implementation of the experiment. In Section 6, the proposed framework is validated and compared on three available datasets and facial videos of licensed crane operators. Section 6 also provides a discussion of the results, and Section 7 concludes the study.

2. Literature Review

Under a fatigued state, crane operators execute the repetitive lift tasks in a complex construction project that may lead to catastrophic casualties as those of the vehicle drivers. There are apparent signs to tell whether an operator/driver is fatigue, such as repeatedly yawning, inability to keep eyes open, swaying the head forward, face complexion changes due to blood flow [8]. As the facial features of the operator/driver in a fatigued state are significantly different from those of the conscious state, the real-time monitoring of the operator/driver's face by the camera can be an efficient, non-invasive and practical approach. In the rest of this section, a review of the tower crane safety issues and existing detection methods for fatigue will be provided.

2.1 Tower Crane Safety

Many research works are being carried out on tower crane safety, including accident analysis, interviews, surveys of construction sites, and modelling analysis [17]. In the operation stage of tower cranes, tower crane accidents are mainly attributed to human factors. According to the analysis of crane accidents occurring in the USA between 1997 and 2003, Beavers et al. [18] found that the low safety performance of the crane operators was the leading cause of crane accidents. In Hong Kong, four major causes of tower crane-related accidents are: (1) fall of persons from height; (2) struck with/by moving objects; (3) struck by falling objects; and (4) trapped by collapsed objects [5]. Tam and Fung [5] also found that operators' unsafe behaviour is the main reason leading to these crane safety issues, especially negligence or misjudgement, inadequate training, multi-level subcontracting systems, schedule pressure, and fatigue of practitioners. Nearly 60.5% of the crane operators were working in a fatigued state due to long working hours with few rests or breaks on demanding physical works. About 52.6% of them were even working without breaks during the whole working day [5]. Shapira and Lyachin [19] conducted structural interviews with surveys and identified that the factors of length of worker shift, operator proficiency, operator character are also essential safety factors influencing tower crane operation. According to accident analysis, inattention or fatigue of the operator is one of the critical causes of tower crane failures [20, 21].

The operations and judgments of the crane operator are thus crucial factors for operational safety and productivity, particularly in the construction site due to the high level of congestion and dynamics [6]. Accordingly, the Construction Industry Council of Hong Kong established

Guidelines on the safety of tower cranes [22], recommended several measures for enhancing tower crane safety, improving site supervision, improving qualification and experience requirements of subcontractors and workers. While the guidelines could only cultivate the safety considerations of practitioners instead of active protections, automatic detection and analysis of crane operator fatigue could provide practical supports for not only crane operators to avoid misoperations but also site superintendents and safety directors to make the proper shift and break arrangement.

2.2 Fatigue Detection Methods

2.2.1 Vision-based fatigue detection

Fatigue is a risk factor at work as it may lead to decreased motivation and vigilance, as well as potential accidents and injuries [23]. With the development of computer vision, more and more fatigue detection algorithms have adopted the technology as underlying learning architecture to analyse the facial features extracted from the video/images. It performs a high accuracy as it facilitates the computer to learn by itself for capturing the key features. For example, Park et al. [24] presented the Driver Drowsiness Detection network consisting of three existing networks by SVMs to classify the categories of videos. However, this approach cannot automatically extract the features of driver drowsiness and monitor driver drowsiness online. Choi et al. [25] trained the hidden Markov models to model the temporal behaviours of head pose and eye-blinking for identifying whether the driver is drowsy or not. These approaches relied on hand-crafted features that have shown limited efficacy in real-time monitoring and can be inaccurate when driver/operator wears the sunglasses or under considerable variation of illumination condition.

2.2.2 CNN-based fatigue detection

In many tasks like image classification and segmentation, object detection, deep learning has achieved notable performances. Therefore, deep learning is considered an effective alternative to evaluate fatigue problems. CNN was first applied to fatigue monitoring as the features extractor of static facial fatigue images by Dwivedi et al. [26]. Then, Zhang et al. [14] used the CNN as both face and nose detectors to show their performances that are quite better than the conventional face detectors such as AdaBoost and WaldBoost with Haar-like features. To achieve real-time fatigue monitoring, Reddy et al. [10] utilized multi-task cascaded CNN with the compression technique to achieve a faster fatigue recognition than existing models of VGG-16 and AlexNet at a reasonable accuracy rate. As described further in [24], various CNN architectures are used to get the feature then classify the fatigue state. They used Alex-Net [27] and VGG-Net [28] as feature extractors. CNN is selected for the generation of a spatial domain feature through a frame-based analysis. Sometimes CNN is also used to handle temporal data, such as 3D CNN, which processes multiple video frames with a specific depth [29, 30]. Concurrently, features learned from unlabelled data based on deep neural networks, such as CNN, have been proved to have a significant advantage over hand-crafted features in real-time monitoring of fatigue [31].

2.2.3 CNN and LSTM-based fatigue detection

Given that the convolution process needs many computing resources, 3D CNN needs to be considered before using it for the real-time scenario. LSTM [32] was proposed as a practical solution for handling the sequences data. It has been proven effective in learning long-term temporal dependencies by solving the exploding and vanishing gradient problems for the traditional recurrent neural network [32]. An LSTM typically comprises a cell and three gates: input, output, and forget. The cell can remember values over arbitrary time intervals, and the gates control the information flow out and into the cell. This is the mechanism considering temporal dependency, and its ability has been tested and widely used in video processing applications [33, 34]. LSTM is also used in some research to evaluate driver drowsiness or human mental workload with EEG and Event-Related Potentials (ERPs) [35, 36].

The integration of CNN and LSTM can be an alternative in fatigue monitoring. Several studies have adopted CNN to extract frame-level features and then feed them into LSTM to extract the temporal features for determining whether fatigue or not. Some refinement techniques help them achieve high accuracy, such as reducing the hidden layer of LSTM [12], noisy smoothing in post-processing [35], alignment technology to learn the most critical fatigue information [37], combine CNN to predict age and to detect the drowsiness in driver and alert them [38].

2.3 Challenges Posed by Datasets

There are numerous works in fatigue detection, but none of them uses a large, public, realistic dataset and is suitable for early fatigue detection of tower crane operators. Due to their constant head moving for tracking the loads' position and recurrent communication (talking) with crane banksman, it is significantly different from the patterns in the available fatigue datasets of vehicle drivers. The primary challenge is to determine which available datasets and the collection methods are most suitable for crane operators' fatigue detection.

On the other hand, it is challenging to compare prior methods and decide what state-of-the-art is in this area [13]. Available fatigue datasets have various collection methods, testing environments, and label principles. It is difficult to compare and evaluate the accuracy of fatigue detection among the available datasets. Furthermore, there is no unified evaluation criterion and labelling principle for the available datasets. Several existing methods [9, 39-42] were evaluated on a small number of datasets without sharing the data sources. In some experiments [10, 43], the participants were instructed to act fatigue instead of obtaining data from subjects who were fatigued. It is then an open question of whether and to what extent videos of pretended fatigue are useful training datasets for detecting real fatigue, especially on early warning purposes.

Therefore, it is a significant benefit to establish a unified evaluation criterion for the available datasets and explore which dataset characteristics and the corresponding data acquisition methods are suitable for crane operators' fatigue detection. Furthermore, the analysis results contribute to building up a large and public realistic fatigue dataset for tower crane operators.

2.4 Research Gaps

The review of related studies reveals three research gaps in fatigue detection of tower crane operators. Firstly, there is a lack of methods that proved to be specifically designed and valuable for tower crane operations. Operators' unsafe behavior is the main reason leading to crane safety issues, especially fatigue causing unsafe practices of tower crane operations. Although fatigue or drowsiness detection is an important research topic and successful cases have been applied in driving or other workplace scenarios, few studies have developed fatigue detection and warning systems for crane operators. The existing methods are difficult to directly use for fatigue detections under crane operation scenarios due to the variations of facial features and different head movement patterns between crane operators and vehicle drivers. Furthermore, the existing methods have usually been tested and specialized on a single dataset, which may not reveal or reflect the variety in crane operations. Secondly, available fatigue datasets have their various collection methods, testing environments, and label principles. It is challenging to compare the general accuracy in fatigue detection among these multi-sources datasets. Thirdly, there is no large, public, and realistic dataset with the subjects on crane operations. As for the fatigue detection of the tower crane operators, the reasonable step is to learn from the existing datasets, explore and analyse the features on such multi-sources datasets, and develop the corresponding data acquisition methods suitable for crane operators' fatigue detection, further providing collection guidelines of crane operators dataset. The existing methods cannot be directly applied to multi-sources datasets. Therefore, it is significant to specially design a method appropriate for the multi-sources datasets to determine which kinds of available dataset’s characteristics and the collection methods are most suitable for crane operators fatigue detection.

This study develops a hybrid learning architecture and comes out with data collection guidelines for tower crane operators' fatigue detection to fill these gaps. It starts by combining CNN and LSTM as the learning architecture. This hybrid learning architecture is adopted for training on three representative and available datasets re-labelled by the authors: NTHU-DDD, UTA-RLDD, and YawDD. Then the trained models are evaluated through licensed crane operators' facial videos. In addition to exploring and analysing which dataset characteristic is suitable for the fatigue dataset for tower crane operators, we also implement a baseline method and include quantitative results from the method in the experiments to show the comparative results accurately.

3.Multi-Sources Fatigue Datasets

In this research, three datasets of vehicle drivers are available for training and evaluating the proposed fatigue detection approach. They are UTA-RLDD, NTHU-DDD, and YawDD, as shown in Fig. 1. Each dataset has its collection method and scenario, label modes, dataset size, and facial expressions on whether the fatigue is "acted" or not. They are used to know the dataset characteristics that are suitable for crane operators' fatigue detection. Further information regarding the three datasets is described as below:

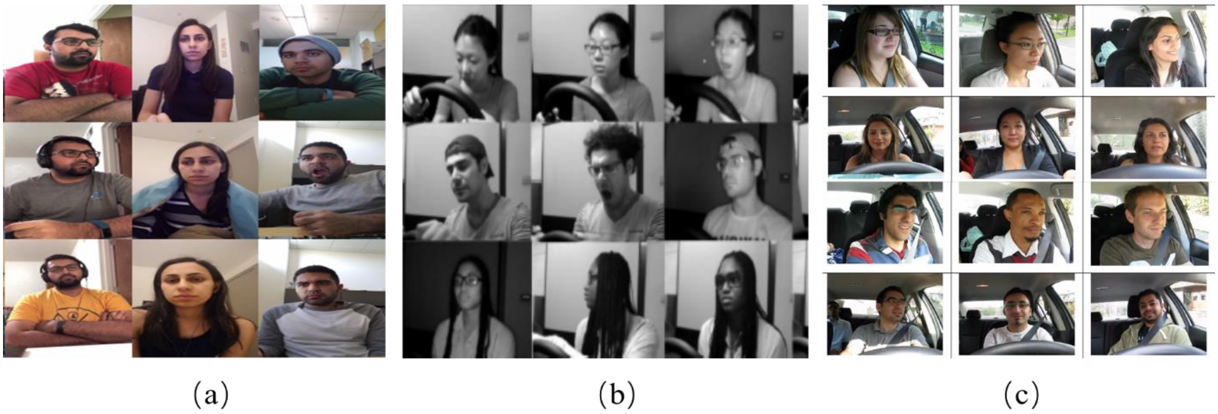


Fig. 1. Multi-sources datasets: (a) UTA-RLDD; (b) NTHU-DDD; and (C) YawDD

The University of Texas at Arlington Real-Life Drowsiness Dataset (UTA-RLDD) [13] was created for the task of multi-stage drowsiness detection. The target of the dataset is focused on discriminative factors like subtle micro-expressions under fatigue cases, not just on extreme and easily observed expressions. Sixty healthy participants recorded 30 hours of RGB videos in the dataset. By using the participant's cell phone or a webcam, they recorded the facial videos by themselves in real life. Therefore, it is expected to detect fatigue or drowsiness at an early stage to activate drowsiness prevention mechanisms through these subtle cases [13]. Participants are difficult to pretend drowsy or fatigue by mimicking subtle micro-expressions because of their physiological and instinctive natures.

The NTHU Driver Drowsiness Detection dataset (NTHU-DDD) [15] is a public dataset collected by the Computer Vision Lab at National Tsing Hua University, which contains 36 IR videos under a variety of simulated driving scenarios. The scenarios include normal driving, yawning, slow blink rate, falling asleep, burst out laughing, and so on. These videos are taken under day and night illumination conditions. However, they are all based on the subjects pretending to be fatigue.

The Yawning Detection Dataset (YawDD) [16] is collected by Distributed and Collaborative Virtual Environment Research Laboratory (DISCOVER Lab) at the University of Ottawa. It contains two available sub-datasets: the first contains 322 RGB videos of normal facial expressions, and the second includes 29 RGB videos targeting driver yawning. Both sub-datasets consist of male and female drivers from different ethnicities with and without glasses/sunglasses. Furthermore, there are three different mouth conditions in the dataset: (1) normal driving with mouth closed (no talking), (2) talking or singing while driving, and (3) yawning while driving. In other applications, it can also be used for yawning and fatigue detection, such as simulating communication between operators and riggers [16].

4. Proposed Fatigue Detection Methods

4.1 Framework

In this study, as shown in Fig. 2 and Fig. 3, we proposed the workflow and architecture of hybrid deep neural networks taking the form of the previous work done by Li et al. [6]. The workflow comprises several steps (Fig. 2). Firstly, capture videos from the field and process the videos to detect the operators' faces. The corresponding landmarks of eyes, mouths, and faces areas are also detected through facial detection. Next, significant fatigue features of the operators contained in these areas are extracted for training fatigue classifiers to realize the fatigue level estimation. Such information can be analysed to see if a warning of fatigue at an early stage is necessary.

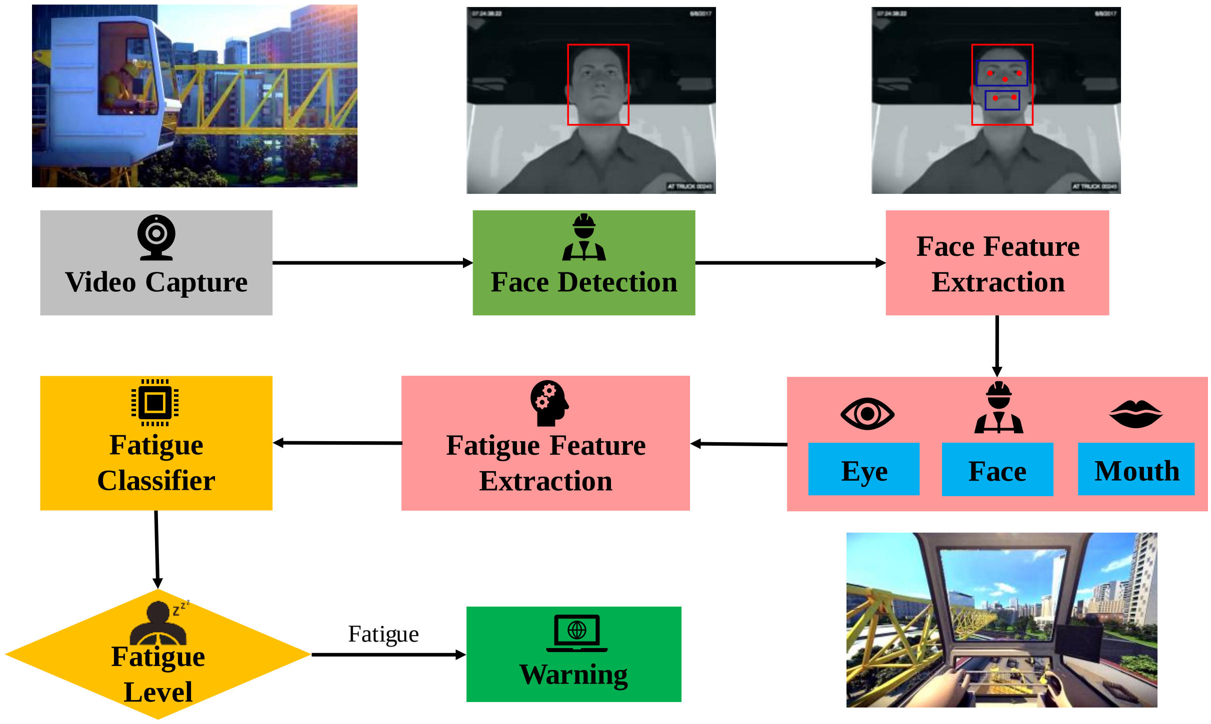


Fig. 2. The workflow of the hybrid deep neural networks for fatigue detection

As shown in Fig. 3, the architecture of the proposed hybrid deep neural networks consists of three main modules: (1) Face Detector, (2) Spatial Feature Extractor, and (3) Temporal Feature Modelling. They are connected through several learning networks. Firstly, the face detector uses Multi-Task cascaded Convolutional Neural Networks (MTCNNs) [44] to allocate the bounding box of the facial area and the corresponding facial landmarks in each frame of the video. The eyes, mouth, and head areas from the facial area are further extracted. Secondly, the customized Efficient Convolutional Neural Networks for Mobile Vision Applications (MobileNet) [45] is adopted as a spatial feature extractor to extract the facial features from the images of the individual frames. Finally, due to the fatigue features follow a pattern over time, an LSTM network is used to leverage the temporal pattern from a sequence of features within a specific time interval. The final output of this architecture is the fatigue level so that a fatigue warning can be further triggered. Each proposed module is detailed in the following sections.

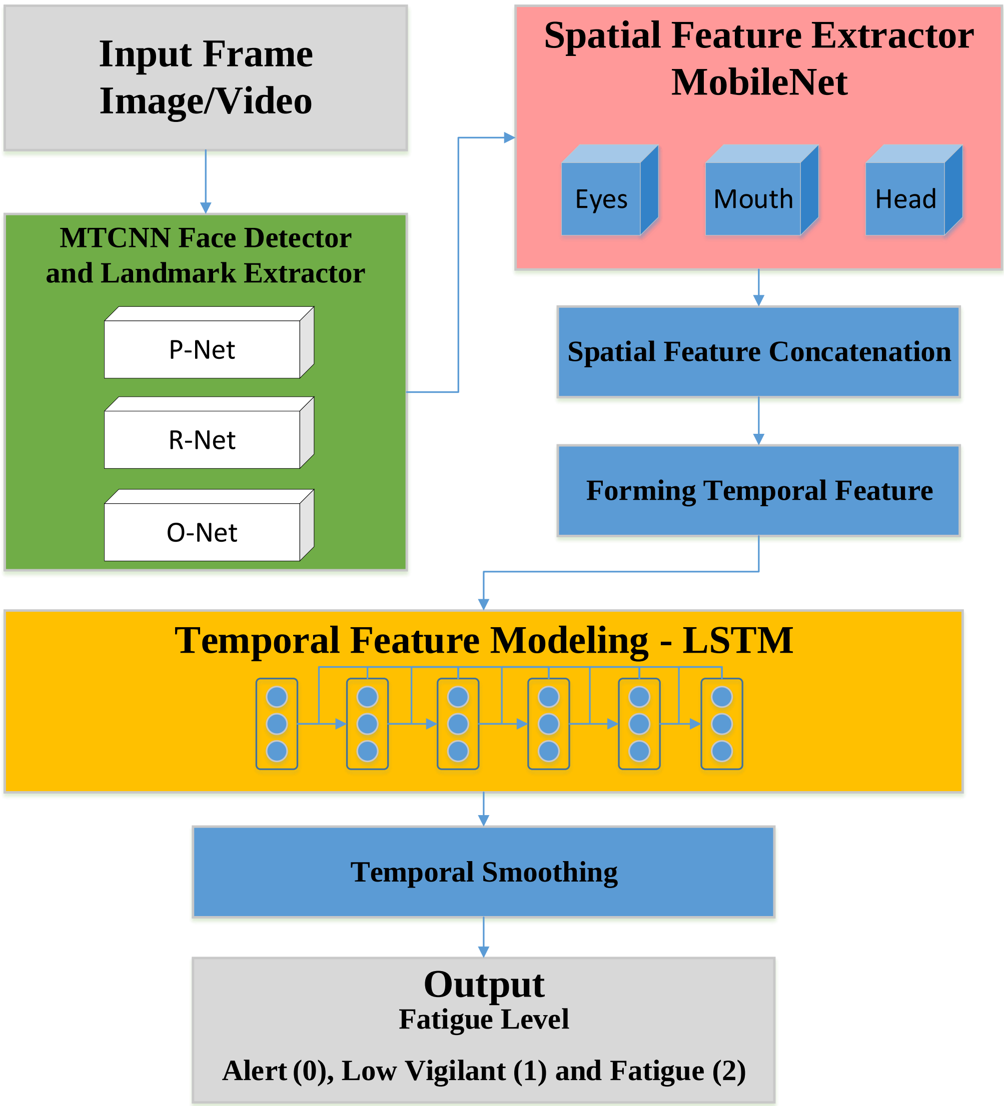


Fig. 3. The architecture of the proposed hybrid deep neural networks for fatigue detection

4.2 Face Detection

Crane operator fatigue detection through videos can be challenging because facial area detection and alignment are affected by many factors, such as the lighting conditions, operator's gestures, video resolutions, facial angles, expressions, and occlusions. Therefore, the design of the face detector is critical to achieving precise facial detection before the facial feature extraction and fatigue detection. The challenges to extract the landmarks of mouth and eye areas could be amplified in crane operation cases because of significant pose variations of the operator. The operator would change pose along with the moving loads, extreme lightings or darkness in the operation cabin, and occlusions between the eyesight and rigging object [6].

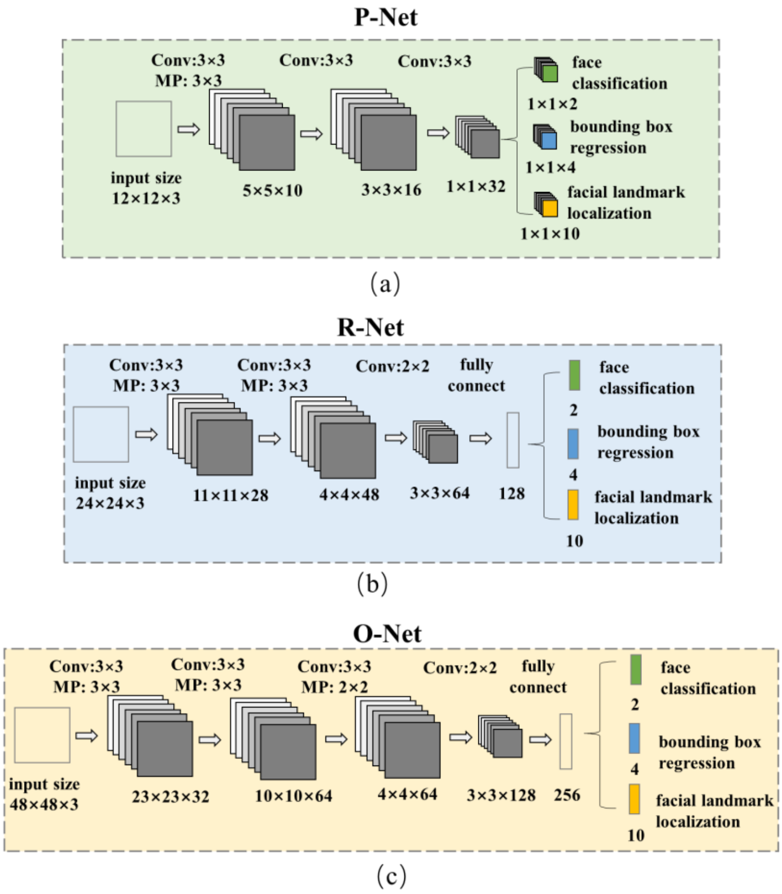


Fig. 4. The architecture of MTCNN: (a) Proposal network (P-Net) structure; (b) Refine Network (R-Net) structure; (c) Output Network (O-Net) structure

The MTCNN proposed by Zhang et al. [44] is known as one of the fastest and most accurate face detectors. In order to solve the challenges mentioned above, MTCNN is applied to conduct the face detection and face alignment tasks with several stages [44]. As shown in Fig. 4, MTCNN consists of three network architectures (P-Net, R-Net, and O-Net) to obtain the facial bounding box and facial landmarks in three different scales.

In MTCNN, $h_{\theta MTCNN}$ is the set of resulting parameters, and I is an input image, as seen in Eq. 1. Through the three networks (P-Net, R-Net, and O-Net), the predicted face bounding box positions are donated as $Sx, Sy, Ex,$ and Ey. Also, the five landmarks, including left eye, right eye, nose, left corner of the mouth, and right corner of the mouth, are donated as $lx0,\ ly0,\ lx1,$ $ly1,\ lx2,\ ly2,\ lx3,\ ly3,\ lx4,\ ly4$. According to the face bounding box positions, the consecutive head areas of the image are cropped as Eq. 2 to get more precise images for the subsequent cropping. Furthermore, based on five landmarks, the eyes and mouth areas are also cropped and extracted by using 30% of the face bounding box size and putting landmarks in the centers, as shown in Eq. 3 and Eq. 4.

$$h_{\theta MTCNN}(I) = [Sx; Sy; Ex; Ey; lx0; ly0; lx1; ly1; lx2; ly2; lx3; ly3; lx4; ly4] \quad (1)$$

$$I_{face} = I[Sx\colon Sy, Ex\colon Ey] \quad (2)$$

$$I_{crop} = I_{face}[xc\colon 0.3w, yc\colon 0.3h] \quad (3)$$

$$\begin{aligned} xc &= x - 0.3w/2 \\ yc &= y - 0.3h/2 \end{aligned} \quad (4)$$

4.3 Spatial Features Extraction

The approach of spatial feature extraction is to learn a CNN-based model for extracting the facial features from the images at individual frames. It includes head, eyes, and mouth determined as the facial landmarks through MTCNN in the face detector. In this study, MobileNet [45] is adopted as the primary approach to enable a fast and stable training process to generate the feature extraction model. The model has achieved good performance in image recognition of various datasets. MobileNet and its variants were introduced as solutions optimized primarily for speed [45]. Fig. 5 demonstrates the improved MobileNet architecture, which includes thirteen convolutional layers (grouped into Conv 1-13), five max-pooling layers (Max Pool 1-5), one average-pooling layer (Ave Pool), and one fully connected feedforward network layer (FC).

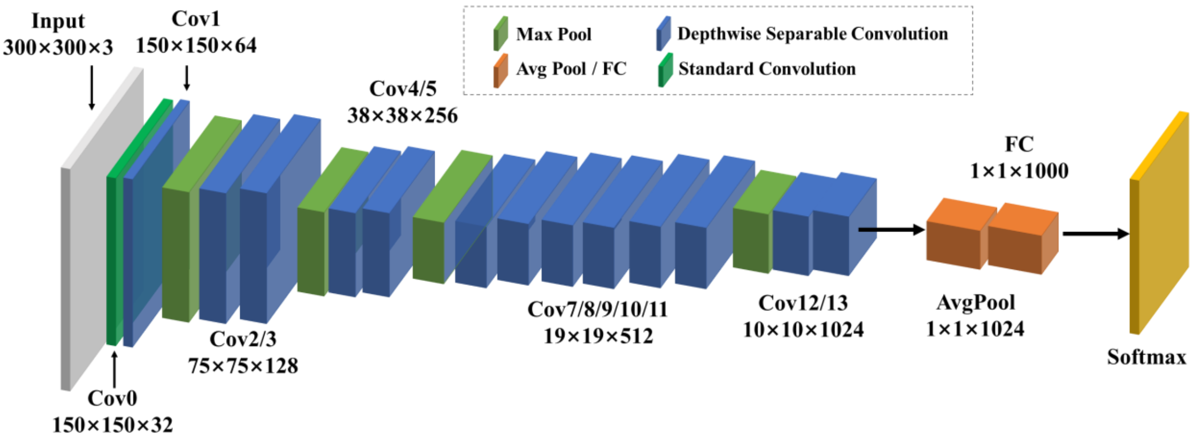


Fig. 5. The architecture of the CNN-based feature extraction model

Fig. 6 illustrates and compares the standard, point-wise, and depth-wise convolutions in the MobileNet. The main building blocks of these classes of networks in this study are depth-wise separable convolutions. The convolution is factorized by two distinct operations: depth-wise convolution and point-wise convolution. It shows that depth-wise separable convolutions have less parameter and computational cost than standard [46].

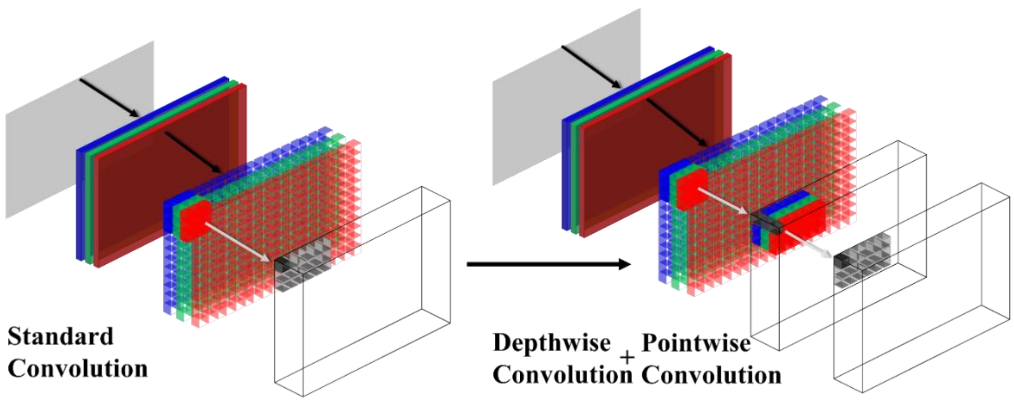


Fig. 6. The infographic of convolution operations

4.4 Temporal Features Extraction

Although the feature extractor can predict the fatigue level of each image frame based on the spatial features, sometimes, it is still hard to discriminate the slight dynamic variations that have strong temporal dependencies, such as yawning and talking. Therefore, it is beneficial to consider temporal information in the sequential frames. To this end, the deep LSTM [47] is applied to model the temporal features. LSTM is a particular type of Recursive Neural Network (RNN) for analysing hidden sequential patterns in both temporal and spatial sequential data [48]. It is capable of learning long-term dependencies because of its unique structure with input, output, and forget gates to control the long-term sequence pattern identification [32]. The LSTM used in the proposed hybrid network is designed to avoid long-term dependency by consisting of gates to control the amount of information that is given during every time frame. The gates work as trying to forget some unimportant information from the previous frame. Meanwhile, it also analyses the information in the current time frame, making an assumption based on the current information and previous important ones.

5. Implementation

In order to identify suitable dataset characteristics and data acquisition methods for crane operators' fatigue detection, three public datasets, named NTHU-DDD, UTA-RLDD, and YawDD, are used as image sources in the proposed hybrid learning architecture. To establish a unified evaluation criterion, training sets of the three datasets are relabelled at the frame and minute segment levels individually. The hybrid fatigue detection models based on the proposed architecture are trained through the training sets. The trained models are further used to evaluate the licensed crane operators' facial videos. The average accuracies and losses are obtained from the validation sets of all datasets (three public ones and the video clips from crane operators) at frame and minute segment levels. Finally, dataset characteristics and the corresponding data acquisition methods for the purpose of crane operators' fatigue detection are discussed. The fatigue detection results are compared and analysed from different perspectives, including human-verified labels at different levels (frame and minute), face poses (front and side view), facial expressions (pretended or real), and illumination conditions.

5.1 Environment Setting

The experiment, including the training and validation process, is conducted in a server running Ubuntu 16.04 operation system. The specification and configuration are as follow:

- CPU: 2× Intel E5-2650v4
- RAM: 8×16GB DDR4 memory
- GPU: 4×GeForce GTX 1080 Ti
- Hard Disk: 240GB SSD and 5 × 4TB HDD
- Run-on CPU: None
- Run-on GPU: MTCNN, MobileNet, and LSTM

For the implementation details, the algorithms are developed using Python with TensorFlow version 1.8.0 and Keras version 2.1.6 as a base deep learning framework. OpenCV 3.3 is also used as an open-source image processing and computer vision library.

5.2 Multi-datasets Descriptions

The experiment consists of videos from two sources: available public datasets, NTHU-DDD, UTA-RLDD, and YawnDD, and videos captured by the authors from interviewing expert operators on performing crane operation simulations in a Unity3D gaming environment. The details of the videos are illustrated in Table 1. They were taken in different scenarios, including working in front of computers, simulated or real driving environments, and simulated crane operations. They have varying facial characteristics, behaviours, ethnicities, illumination conditions, acquisition scenarios, and face poses (different camera positions). Videos are also captured with different resolutions, such as 640 × 480, 1280 × 720, and so on.

Table 1. The detailed information on the available datasets

Dataset	Subjects	Behaviour	Illumination	Camera Type	Scenarios	Age	Camera Position
NTHU-DDD	36	• Stillness • Yawning • Nodding • Looking aside • Talking and laughing • Sleepy eyes • Drowsy	Day and Night	Active Infrared (IR)	• Bare face at daytime • Glasses at daytime • Sunglasses • Bare face at night • Glasses at night	-	On the top of the screen of the laptop or mobile phone (Similar to the place on the vehicle dashboard)
UTA-RLDD	60	• Alert • Low Vigilant • Drowsy	Morning Noon Midnight	RGB	• Glasses • Sunglasses • Moustache • Bread • Bare Face	20-59	On the vehicle dashboard
YawnDD	107	• Normal • Talking • Yawning • Singing	Day (from early morning till sunset)	RGB	• Glasses • Sunglasses • Moustache • Bread • Bare Face	-	Under the front vehicle windshield On the vehicle dashboard
Licensed Crane Operators	5	• Alert • Low Vigilant • Fatigue	Day	RGB	• Glasses • Hat • Bare Face	30-50	On the top of the screen of the laptop (Similar to the place on the vehicle dashboard)

5.3 Multi-datasets Relabelling

During the experiment, a problem comes due to the long-term dependency in a specific dataset, which is that those alert facial expressions on a series of frames, within a few seconds, would still be considered as drowsy signs if they had just restored the expressions to alert after drowsy states [37]. Furthermore, the level of details on existing labels in this dataset cannot identify the drowsy states with high precision in the temporal dimension [37]. Compared with other datasets, it also shows no unified evaluation criterion and labelling principles among them. To address these problems, the authors relabelled the three available datasets NTHU-DDD, UTA-RLDD, and YawnDD, with every frame and minute as segment units. Those typical facial states or behaviours, such as closing eyes, yawning, and lowering head, are still considered as the evidence to judge whether a frame contributes to the awareness of fatigue. To describe the transitional states between the alert and the fatigue, as well as establish a unified evaluation criterion, we propose a relabelling workflow and a unified relabelling principle.

5.3.1 Datasets relabelling workflow

The proposed workflow of datasets relabelling consists of the following steps: (1) video transform; (2) state and behaviour description; and (3) fatigue level labelling. The videos are transformed into a sequence of images by Python script to relabel the available datasets at the frame segment level. According to frame indexes stored as CSV files, different facial states and behaviours are manually labelled in each frame. Their specific state and behaviour are recorded manually. After the state and behaviour description, they are further transformed into three fatigue levels, alert, low vigilant, and fatigue, for further fatigue classifier training [13].

5.3.2 Relabelling principles

Karolinska Sleepiness Scale (KSS) [11] is adopted and modified in this relabelling process to describe the fatigue level. KSS is a 9-point Likert scale. The KSS scores are defined in Table 2 (1 - Extremely alert; 3 – Alert; 5 - Neither alert nor sleepy; 7 - Sleepy, but no difficulty remaining awake; and 9 - Extremely sleepy, fighting sleep). The scale is often used when studies involving self-reporting and subjective assessment of an individual's drowsiness at the time. Scores of KSS usually increase with longer periods of wakefulness, and they strongly correlate with the time in a day.

Table 2. Karolinska Sleepiness Scale (KSS)

Karolinska Sleepiness Scale (KSS)	
Extremely alert	1
Very alert	2
Alert	3
Rather alert	4
Neither alert nor sleepy	5
Some signs of sleepiness	6
Sleepy, but no difficulty remaining awake	7
Sleepy, some effort to keep alert	8
Extremely sleepy, fighting sleep	9

For establishing a unified relabelled principle, a modified approach is developed for multi-datasets learned from the KSS and the labelling approach of UTA-RLDD. The fatigue states are classified into three different levels, which are defined as follows:

1) Alert (labelled as 0): The top four levels (1 - Extremely alert; 2 - Very alert; 3 – Alert; and 4 - Rather alert) in the KSS are merged as alert, which means the subject is experiencing no signs of sleepiness.
2) Low Vigilant (labelled as 1): As stated in levels 5 and 6 (Neither alert nor sleepy and Some signs of sleepiness) of KSS, low vigilant corresponds to subtle cases when some signs of sleepiness appear, or sleepiness is present, but no effort to keep alert is required.
3) Fatigue (labelled as 2): The levels 7, 8, and 9 (Sleepy, but no difficulty remaining awake, Sleepy, some effort to keep alert and Extremely sleepy, fighting sleep) in the KSS are categorized as fatigue, which means the subject needs to try not to fall asleep actively.

In order to make the sequential labels reflect the practical driving environments, the unified relabelling principles are put forward in more human-verified states and behaviours according to the integration of the three public dataset’s behaviour descriptions. As shown in Table 3, many states consisting of driver expected behaviours, states changing from alert to fatigue, or some subtle signs of fatigue, are classified into the three fatigue levels individually. Apparently, the behaviours, such as stillness, looking aside, normal blinking and talking, laughing, and singing, are the least related to fatigue. Therefore, they can be relabelled to 0. In order to achieve early fatigue detection, behaviours, like distraction and sleepy blinking are defined as the states standing for the changes from alert to fatigue or some signs of fatigue. They can be relabelled to 1. As for obvious fatigue behaviours, like yawning and nodding, they can be relabelled to 2. The relabelling takes into account and emphasizes more on the transition from alertness to fatigue.

Table 3. The proposed relabelling principles for multi-datasets

Behaviour	**Description**	**State**	**Fatigue Level**
Talking, laughing, or singing	The driver is talking or laughing while driving		
Looking aside	The driver turns his head left and/or right	Alert	0
Normal blinking	The driver is normally blinking		
Stillness	The driver drives normally		
Distraction	The driver loses focus during driving	Low Vigilant: transitional states between alert and fatigue	1
Sleepy eyes	The driver closes his/her eyes due to drowsiness while driving		
Sleepy blinking	The driver is slowly blinking		
Drowsy	The driver looks sleepy and lethargic		
Yawning	The driver opens his/her mouth widely due to tiredness	Fatigue	2
Nodding	The driver's head falls forward when drowsy or asleep		

An example of the relabelled sequential frames is illustrated in Fig. 7. After the videos are transformed into a sequence of image frames, each frame is described by specific behaviour, like alert, sleepy-eyes, yawning, as shown in Table 3. The descriptions are further transformed into three fatigue levels for further fatigue detection training process. The same approach is adopted to relabel the frame of video clips at every minute.

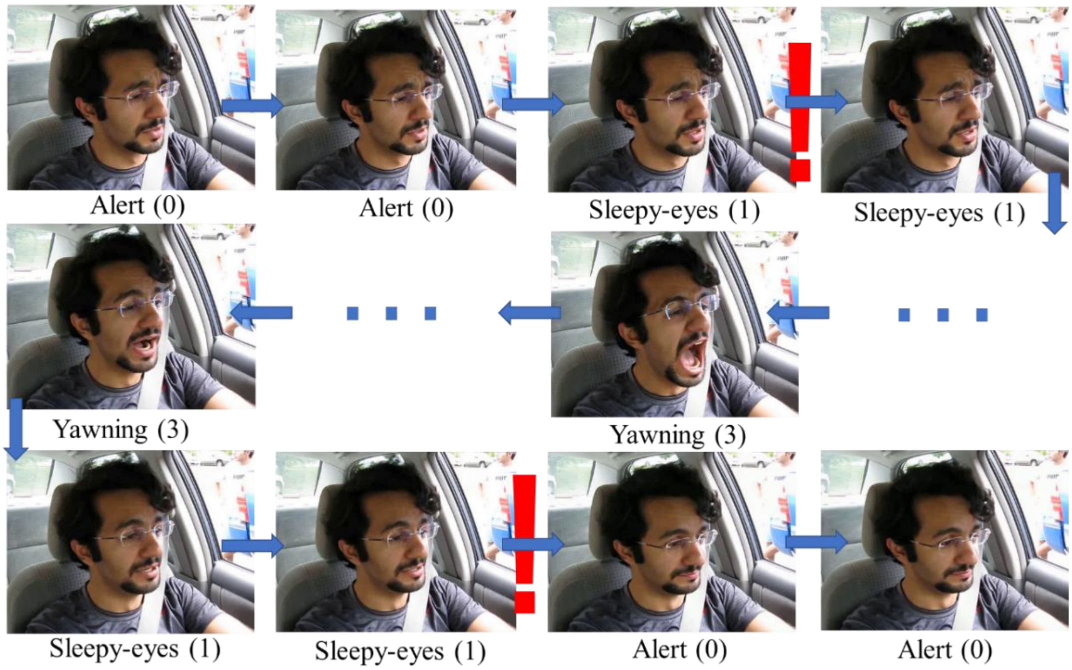


Fig. 7. An example of the relabeled sequential frames

5.4 Data Pre-processing

For all videos, the MTCNN is used to detect the faces from all frames. The detected face bounding boxes with five landmark points are cropped along with boundary pixels, and the cropped face regions are resized to a fixed size 64 × 64. As it is time-consuming in line with the high frame rates of the dataset (e.g., 30 fps or 15 fps), this study sub-sampled the video frames by the factors of 6 or 3 and input the face sequence in a frame rate of 5 fps to the proposed hybrid neural network. The classification results (predicated levels) can be up-sampled back to the original video length. In addition, some videos in the datasets are grey-scale ones. Thus, each frame should be replicated three times to become 3-channels images so as to generalize the proposed method in processing either colour or grey-scale inputs.

5.5 Training and Test Procedure

In the experiment, the two classifiers of the proposed learning architecture, referred to as Spatial Feature Extractor (MobileNet) and Temporal Feature Modelling (LSTM), are trained and evaluated separately. Then the entire architecture is tested by combining the two trained classifiers. The general training and test produce are detailed as follows:

1) All videos in the three public datasets were trimmed into video clips with a fixed length that can start from an arbitrary frame of the original video. The sequential features computed from humans' eyes, mouths, and head areas in one video clip were considered. From all available data obtained from video clips, 70% were randomly selected to train the classifiers, which was then evaluated using the other 30% from the remaining data.
2) For the Spatial Feature Extractor (MobileNet), the eyes, mouths, heads areas detected from customized MTCNN with three fatigue levels: alert, low vigilant, and fatigue, were used for training and validation. Furthermore, this customized MobileNet was used for extracting sequential features for further training at the next step.
3) For the Temporal Feature Modelling (LSTM), it was used to leverage the temporal pattern from a sequence of features within a specific time interval due to the fatigue features follow a pattern over time. It was also trained and evaluated through the same set of randomly selected 70% and 30% data from all available datasets.
4) After training the two classifiers, specifically for the three datasets individually, the data were selected to test the trained models with the integration of the two classifiers. Then all the final performance was evaluated against the ground truth labels.

5.6 Evaluation Metrics

The performance of the proposed fatigue detection architecture on multi-datasets was evaluated quantitatively in terms of accuracy and loss. To achieve more detailed and precious fatigue level prediction, the Mean Absolute Error (MAE) is used as the loss metric because the fatigue level detection is a multi-class ordinal classification problem. It can be considered an intermediate problem between regression and classification [49]. Furthermore, as for the multi-class ordinal classification, Gaudette and Japkowicz [50] compared various metrics for ordinal classification accuracy, and they showed that, as a single statistic, the MAE (Mean Absolute Error) or MSE (Mean Squared Error) performed better than the other measures that they found in the literature. Although MAE/MSE is designed for continuous data, its property of penalizing deviations from the mean more severely works well for ordinal data converted to small integers.

The evaluation metrics are defined in Eq. 5 and 6. Accuracy is the primary metric in this study. It refers to the percentage of entire videos, not individual video clips, that have been classified correctly:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

TP, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative individually, based on the comparisons between the fatigue detection results and ground truths.

Loss (i.e., Mean Absolute Error, MAE) is the average absolute difference between the estimated value and the actual value:

$$Loss = \sum_{i=1}^{N} \left| Y_i - \widehat{Y_i} \right| / N \tag{6}$$

Y_i denotes the fatigue level being predicted and $\widehat{Y_i}$ represents the actual label value. N is the number of video frames being used for fatigue detection.

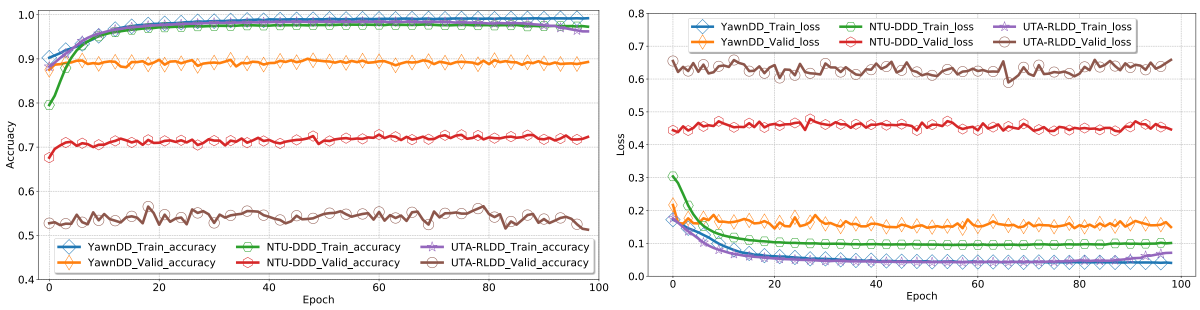
6. *Experimental Results and Discussion*

To determine the suitable dataset characteristics and acquisition scenarios for crane operators' fatigue detection, the processed videos with fatigue levels from four datasets are analysed, as shown in Fig. 8. The experiment results, including the performance of the proposed architecture, influence of segment details, camera positions, illumination conditions, and validation under simulated crane operation scenarios, are described in the following sub-sections:

Fig. 8. Fatigue detection results from multi-datasets: (a) NTHU-DDD; (b) YawnDD; and (c) UTA-RLDD

6.1 Performance of Proposed Architecture on the Available Multi-Datasets

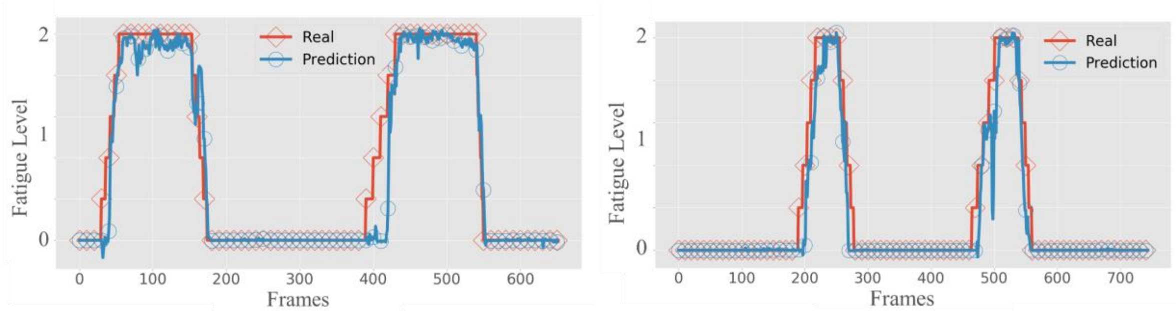
Fig. 9 shows the overall accuracy and loss of the training sets and the validation sets of the three available datasets. The proposed hybrid deep neural network architecture with CNN and LSTM achieves 54.71%, 72.76%, and 87.52% accuracy on the validation sets of UTA-RLDD, NTHU-DDD, and YawnDD individually (with models trained through labels at every frame). The overall accuracies for all sets increase and the overall losses decrease along with the training epochs growing. Because the features fed into the LSTM are extracted by the Spatial Features Extraction (MobileNet), the LSTM has achieved a certain performance of the fatigue detection at the beginning of the training. The results show that the fine-tuning process of the learning architecture specialized for each dataset is unavoidable if the training requires to maximize the detection performance. Nevertheless, the proposed hybrid deep neural network set a baseline for the comparisons to extract significant features for effective fatigue detections.



(a) (b)

Fig. 9. Performance of the proposed architecture on the three public datasets: (a) accuracy and (b) loss

In some cases, on fatigue level detection, the predicted fatigue levels generated by the trained models align well with original labels, as shown in Fig. 10. It shows that the individual trained models matched the fatigue signs patterns among each dataset, and training processes are conducted effectively.



(a) (b)

Fig. 10. Comparison between the predicted fatigue level and original label

6.2 Influence of Windows Size and Layer Configuration of LSTM

Due to the randomness in the neural network's training and verification process, the evaluation results are slightly different depending on the learning architecture configurations compared to the results in the previous section. Table 4 shows the performance of the proposed architecture with different input window sizes on the datasets NTHU-DDD and UTA-RLDD. The two datasets contain video frames with complex behaviors or subtle fatigue facial features. If the window size of LSTM increases, then the accuracies of the proposed architecture increase significantly, which in turn slowdowns the training. Contrarily, input window size decreases, making the accuracy decreases, and the network getting trained faster. While for the YawDD dataset with apparent fatigue facial features, there was no significant increment in the performance of the proposed architecture. Therefore, if the performance should be improved, expanding the window sizes of LSTM can be the right choice for considering more features in dealing with longer time-series data. Besides, there should be a balance between the performance and the training cost of the architecture.

Table 4. Performance of proposed architecture with different window sizes

Training Dataset	Facial Expression	Segment Level	Window Size (pixel)	Accuracy	Loss
NTHU-DDD	Pretend	Frame	15	0.5513	0.6763
			30	0.6073	0.6007
			45	0.7481	0.4191
			60	0.7212	0.4786
YawDD	Pretend	Frame	15	0.8347	0.2295
			30	0.8437	0.2179
			45	0.8444	0.2232
			60	0.8488	0.2137
UTA-RLDD	Real	Frame	15	0.4158	0.8019
			30	0.5325	0.7169
			45	0.5903	0.5884
			60	0.6448	0.5305

Table 5 shows the performance of the proposed architecture with different LSTM layer configurations. Among the three datasets NTHU-DDD, YawDD, and UTA-RLDD, simply increase or decrease the number of layers in LSTM cannot effectively affect the performance. To sum up, for improving fatigue detection performance, expanding the window sizes of LSTM would have apparent effects, especially for the datasets with complex behaviors or subtle fatigue facial features that are more difficult to identify.

Table 5. Performance of proposed architecture with different layers of LSTM

Training Dataset	Facial Expression	Segment Level	LSTM Structure	Accuracy	Loss
NTHU-DDD	Pretend	Frame	512(LSTM) *128(Dense)	0.7225	0.4536
			512(LSTM)*256(LSTM) *128(Dense)	0.7327	0.4582
			512(LSTM)*256(LSTM)*128(LSTM) *128(Dense)	0.7212	0.4429
YawDD	Pretend	Frame	512(LSTM) *128(Dense)	0.8728	0.2298
			512(LSTM)*256(LSTM) *128(Dense)	0.8752	0.2288
			512(LSTM)*256(LSTM)*128(LSTM) *128(Dense)	0.8781	0.2193
Real	Real	Frame	512(LSTM) *128(Dense)	0.5139	0.6602
			512(LSTM)*256(LSTM) *128(Dense)	0.5325	0.6130
			512(LSTM)*256(LSTM)*128(LSTM) *128(Dense)	0.5229	0.6305

6.3 Influence of Label Segment Levels and Real (or Pretended) Facial Expression

Table 6 represents the average losses and accuracies on the three available datasets under different facial expression approaches and label segment levels. In terms of accuracy affected by the trained models of different segment levels, the trained models generally work better on NTHU-DDD (72.76% on the labels with frame segment level and 67.54% on those with minute segment level) and YawDD (87.52% on frame segment cases and 72.63% on minute segment cases). Both datasets contain the pretend facial expression cases in actual or simulated driver environments. However, the trained models work less effectively on UTA-RLDD (54.71% on frame segment cases and 48.05% on minute segment cases). This dataset contains subtle fatigue facial features in daily life environments. The subtle fatigue facial expressions show fewer apparent features to be captured through the training process; thus, it can be the potential reason given the much lower accuracy obtained.

As for segment levels, the results on the three datasets all suggest that labels with frame segment levels come out with better training and prediction performance than those with minute segment levels. However, at the frame segment level, the labelling process generally and naturally takes much more time (around 60 times on average). The effort to put labels at every minute segment can still be considered to save cost and time. As a trade-off, the training model with the datasets labelled at the minute segment level can achieve relatively lower accuracy.

Table 6. Performance of proposed architecture with different segment levels and facial expression approaches

Training Dataset	Facial Expression	Validating Dataset	Facial Expression	Segment Level	Loss	Accuracy	Label Time Spent (min)
UTA-RLDD	Real	UTA-RLDD	Real	Frame	0.6529	0.5471	3720
				Minute	0.7763	0.4305	62
NTHU-DDD	Pretend	NTHU-DDD	Pretend	Frame	0.4556	0.7276	2480
				Minute	0.4860	0.6754	41
YawDD	Pretend	YawDD	Pretend	Frame	0.2288	0.8752	2970
				Minute	0.4256	0.7263	49.5

The proposed trained models (trained by labels at frame level) are used for a cross-checking process in the evaluation to further explore the facial expressions' influence. The three corresponding models are tested on the other two datasets to compare the model applicability. The results can be seen in Table 7; the accuracies are between 30% and 80% for the trained models from one dataset to testing on the other two datasets. It is worth mentioning that the accuracy of the trained model through UTA-RLDD and testing through YawDD is up to 80.27%, which is higher than that on the original evaluation set of UTA-RLDD. The results indicate that although the videos in UTA-RLDD contain real and subtle facial expression cases under the real environment, it is relatively challenging to identify fatigue levels. It surprisingly achieves higher accuracy on videos of the other two testing datasets (YawDD and NTHU-DDD) with pretend and obvious fatigue expressions. On the contrary, the trained models on the datasets with pretend and obvious fatigue expressions (YawDD and NTHU-DDD) have lower accuracies on testing the other datasets. It shows that subtle facial expression captured under real operation scenario is still necessary because the subtle facial features are more sensitive to detect obvious signs of fatigue. It will be the key for early fatigue detection given that another way around, using obvious facial features to detect subtle fatigue signs is less effective.

Table 7. Performance of trained models testing on the other two datasets

Training Datasets	Facial Expression	Segment Level	Testing datasets	Facial Expression	Segment Level	Loss	Accuracy
UTA-RLDD	Real	Frame	NTHU-DDD	Pretend	Frame	1.3338	0.5113
			YawDD	Pretend		0.4383	0.8027
NTHU-DDD	Pretend	Frame	UTA-RLDD	Pretend	Frame	1.5226	0.3528
			YawDD	Real		1.0841	0.4423
YawDD	Pretend	Frame	UTA-RLDD	Real	Frame	1.1087	0.5686
			NTHU-DDD	Pretend		1.4204	0.5342

6.3 Influence of Camera Positions

The YawDD dataset, containing videos captured by different camera positions, is selected for the testing to determine the appropriate facial video capturing angles. As shown in Fig. 11, YawDD contains two video sets of drivers with various facial features for yawning detection. In the first set, a camera is mounted under the front vehicle windshield with an angle to face to the driver (side view). The camera is mounted on the vehicle dashboard in the second set, directly facing the drivers (front view). In the dataset, each driver has three or four videos. Each video contains facial expressions with different mouth conditions, such as stillness, talking/singing, and yawning. This dataset provides 322 videos consisting of both male and female drivers, with and without glasses/sunglasses, different ethnicities, and under three different scenarios: (1) normal driving (no talking); (2) talking or singing while driving; and (3) yawning while driving.

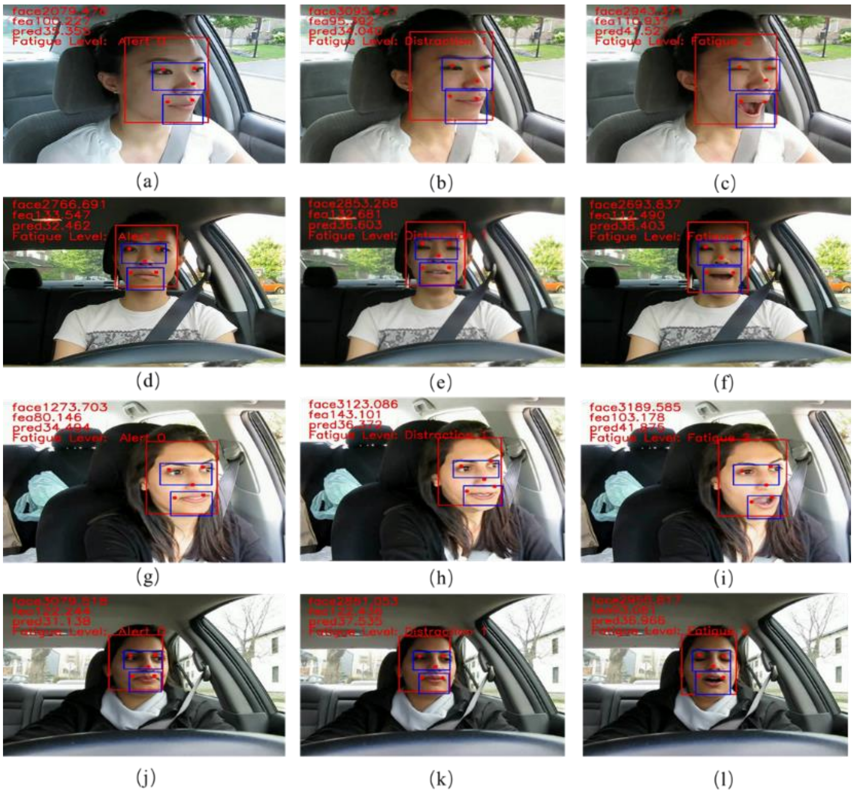


Fig. 11. Facial videos captured by cameras with different positions: (a)-(c) and (g)-(i) are side views; and (d)-(f) and (j)-(l) are front views

As shown in Table 8, the accuracy of the fatigue detection from the videos captured in the driver's front view is 89.34%, which is higher than the accuracy of 82.23% comes out from the videos with the side view drivers. It fits the natural expectation that the more the face portion to be captured, the more features to be detected to increase fatigue detection accuracy. Though, the detection performance of the trained model through side-view videos still achieved high accuracy in this experiment.

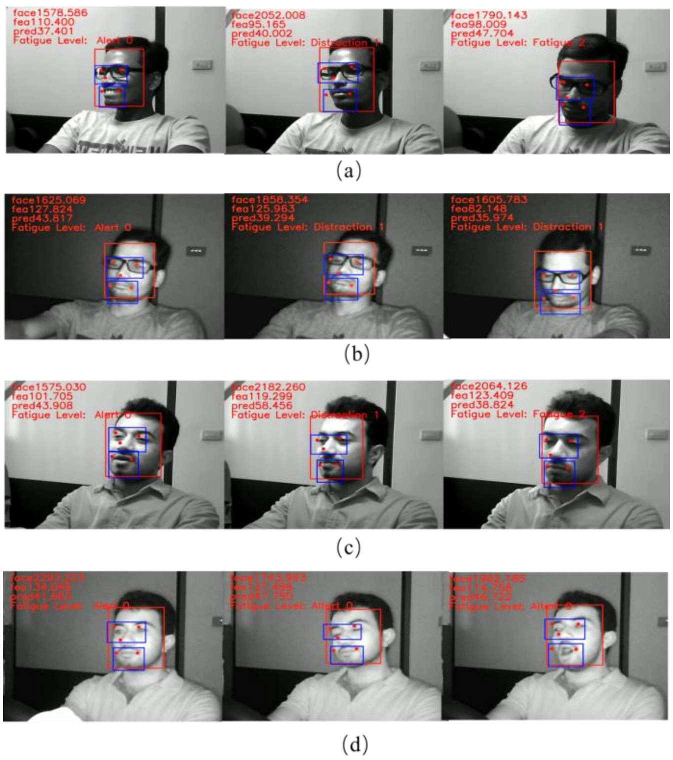
Table 8. Performance with different camera positions

Datasets	Camera Positions	Loss	Accuracy
YawDD	On the vehicle dashboard (front view)	0.2288	0.8934

	Under the front vehicle windshield with an angle to face to the driver (side view)	0.4068	0.8232

6.4 Influence of Illumination Conditions

The NTHU-DDD dataset, containing the same person's videos from different illumination conditions at daytime and night, is selected for comparison to determine the influence of complicated illuminations. An IR camera captures the videos in the daytime and night. The performance of the proposed architecture works well for the videos in the daytime (as shown in Fig. 12(a) and (c)), while, in some cases, it may fail to detect the correct fatigue level at night (Fig. 12 (b) and (d)). The features of fatigue extracted through the MobineNet are inconsistent in the numerical distribution, leading to the wrong classification of fatigue. Lacking the illuminations to enhance the different intensities on the regions of the face may cause these detection challenges. Nevertheless, using IR cameras for facial video capturing still leads to successful detections in many cases.



(a) (b) (c) (d)

Fig. 12. Fatigue detection for illumination conditions at: (a) and (c) daytime; and (b) and (d) night

6.5 Validation on Crane Operators in Simulated Crane Operation Scenarios

Due to a lack of facial videos on real tower crane operators under the operations, the authors invited licensed operators to validate the fatigue detection models through their performance under a simulated crane operation environment, as shown in Fig. 13. It is designed to capture facial videos through a webcam for fatigue detection when operating a virtual crane by following rigging instructions. All the data, including portrait rights, are authorized by the interviewed operators for academic studies and publications. The frames of the videos are labeled into three fatigue levels (alert, low vigilant, and fatigue) as well according to the proposed relabeling principles of multi-datasets. Totally five operators have been invited to participate in the simulation and data collection process. The three trained hybrid learning models are used to detect fatigue from the captured videos, as shown in Fig. 14. These results show opportunities to explore which dataset characteristics are more suitable to be considered in future data collection under operator fatigue detection scenarios.

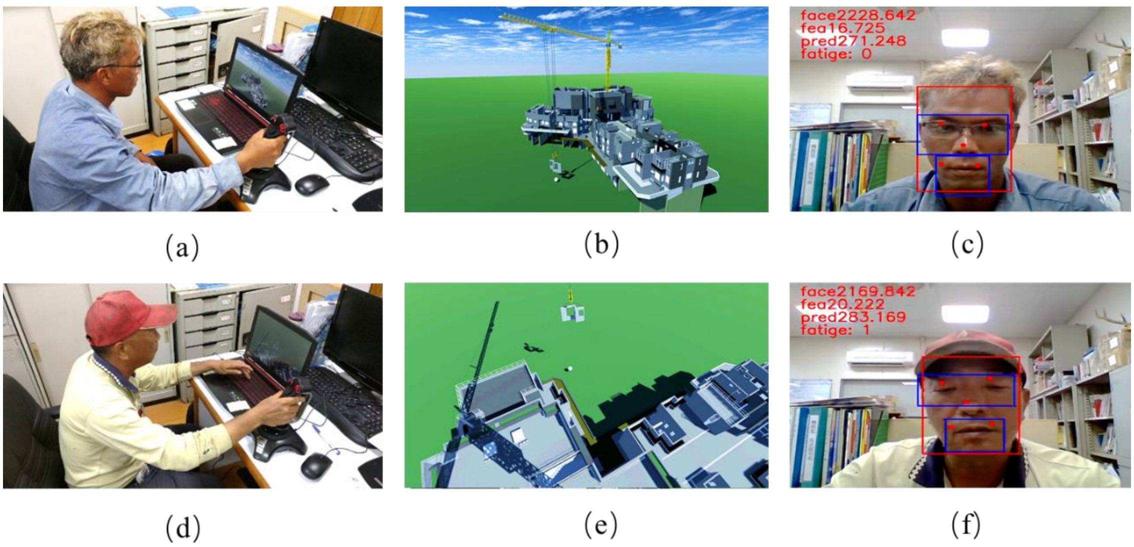


Fig. 13. Simulated crane operations performed by licensed operators: (a) and (d) overview; (b) and (e) tower crane simulation; and (c) and (f) fatigue detection through facial features

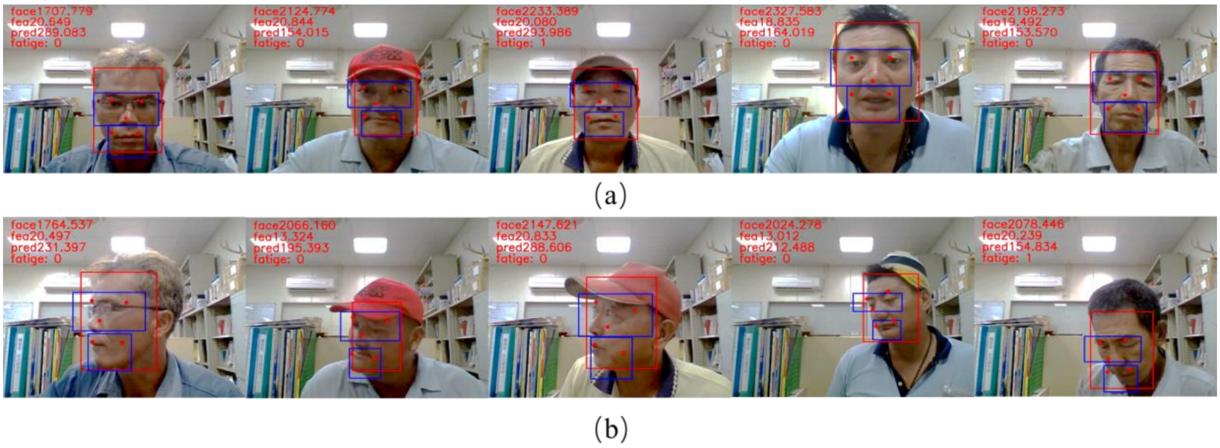


Fig. 14. Fatigue detection results on facial videos of the five crane operators: (a) front faces; and (b) partial faces

Table 9 and Fig. 15 show the average accuracies and losses of fatigue detection on the crane operators' facial videos. The average loss by using the trained model of YawDD is 0.1602, which is lower than those of NTHU-DDD (1.9983) and UTA-RLDD (0.2378). Also, the average accuracies are 78.28%, 29.96%, and 92.81% by using the trained models of UTA-RLDD, NTHU-DDD, and YawDD individually. In general, the trained model from YawDD with obvious facial features achieved the best detection accuracy and lowest loss. At the same time, that of UTA-RLDD with subtle facial features also came out with relatively better results. The bias could cause the lower accuracy and higher loss of the model trained by NTHU-DDD, given that the training on some videos under low illumination conditions in this dataset. It suggests that separated training processes under different illumination conditions (daytime and night) or more videos to be collected for training and evaluation may be needed.

Table 9. Performance of the three trained models on the crane operators' facial videos

Dataset	Performance	Operator1	Operator2	Operator3	Operator4	Operator5	Average
UTA-RLDD	Loss	0.1953	0.3266	0.1769	0.0975	0.3962	0.2378
	Accuracy	0.8599	0.6757	0.8763	0.9475	0.5886	0.7828
NTHU-DDD	Loss	2.2918	2.1798	0.2194	3.2651	0.7913	1.9983
	Accuracy	0.1678	0.1548	0.6938	0.0000	0.3582	0.2996
YawDD	Loss	0.1841	0.0014	0.2797	0.0022	0.3343	0.1602
	Accuracy	0.9164	1.0000	0.9006	1.0000	0.8266	0.9281

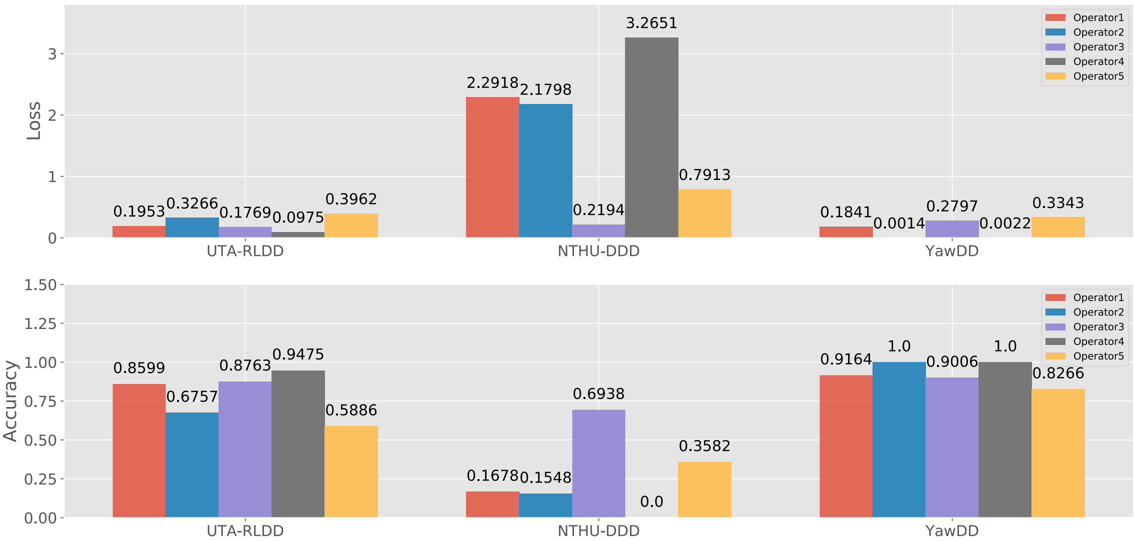


Fig. 15. Histogram of the performance on the crane operators' facial videos through the three trained models

To further determine whether there are significant differences among the detection performance of the three trained models under crane operation scenarios, the paired t-tests are used to identify the significance in terms of the loss and accuracy (Table 10). As for the loss, the *p*-value between the trained model performance of NTHU-DDD and UTA-RLDD, and NTHU-DDD and YawDD are 0.058 and 0.059 individually, which means deviations between NTHU-DDD and another two datasets are significant. Similarly, the accuracy results show substantial differences between NTHU-DDD and UTA-RLDD, and NTHU-DDD and YawDD, given that their *p*-values are both less than 0.05. They suggest that the trained model from NTHU-DDD did result in less effective performance under operator fatigue detection scenarios.

While the results show that the accuracy and loss between UTA-RLDD and YawDD have no significant difference in terms of performance, it further confirmed the importance of subtle facial features (characteristic of UTA-RLDD) under the crane operation scenarios.

Table 10. Paired t-test based on the fatigue detection results of the trained models on the crane operators' facial videos

T-test	Datasets	Mean ± Standard Deviation	Datasets	Mean ± Standard Deviation	t	p
Loss	UTA-RLDD	0.24±0.12	NTHU-DDD	1.75±1.23	-2.629	0.058
	UTA-RLDD	0.24±0.12	YawDD	0.16±0.15	1.112	0.328
	NTHU-DDD	1.75±1.23	YawDD	0.16±0.15	2.609	0.059
Accuracy	UTA-RLDD	0.79±0.15	NTHU-DDD	0.27±0.27	3.595	0.023
	UTA-RLDD	0.79±0.15	YawDD	0.93±0.07	-2.325	0.081
	NTHU-DDD	0.27±0.27	YawDD	0.93±0.07	-4.625	0.010

In summary, the experiment results indicate that the proposed learning architecture works with effectiveness on the crane operators' fatigue detection. Among the available datasets, the dataset with apparent fatigue facial features in actual or simulated driving environments is comparatively easier for detection than those with subtle fatigue facial features in indoor environments. However, the subtle fatigue facial features are still contributing to accuracy positively. Also, labelling resolution significantly affects detection. The trained model performance from the human-verified labels at the frame segment level is more accurate than those with a minute segment level for detecting operators' fatigue. As for the variation of face pose, the videos with side view facial expressions are more difficult to detect the subject's fatigue accurately than those captured through the front view. In order to avoid the influence of complicated illuminations, the IR camera can be used along with the RGB camera for the scenarios at night and train the separated models under different illumination conditions (daytime and night). Still, the comparisons of the experiments show that the detection of videos at daytime is more accurate than those captured at night by the IR camera.

7. Conclusion and Future Work

This study identifies and discusses the guidelines for collecting crane operators' facial videos for fatigue detection during operations. A hybrid learning architecture as a unified evaluation criterion is proposed by combining CNN and LSTM to detect the fatigue status based on three public datasets, NTHU-DDD, UTA-RLDD, and YawnDD, with vehicle drivers' facial videos. In order to identify the necessary dataset’s characteristics and suitable data collection approaches for crane operators' fatigue detection, the comparative experiments are conducted on the three public datasets and tested on the facial videos of crane operators through a simulated crane operation environment. The preparation of the experiments includes relabelling video clips with the segment level at every frame and minute and used the proposed learning architecture to train the hybrid fatigue detection models based on the three datasets separately.

The contributions of this study are fourfold: (1) expand the fatigue detection approaches from vehicle drivers to crane operators; (2) the trained hybrid learning models showed its feasibility to uniformly detecting the facial regions with critical fatigue features; (3) the exploration and analysis on which dataset characteristics and the corresponding data collection approaches are suitable under crane operators' fatigue detection scenario; and (4) give guidance for building up a large and public realistic fatigue dataset for crane operators.

Based on the study results, there are suggestions for establishing a large and public fatigue dataset for tower crane operators: (1) The datasets with apparent fatigue facial features under real driving scenarios are comparatively accurate for the detection than those with subtle fatigue facial features. However, the trained model from subtle fatigue facial features has achieved equal accuracy on the operator's fatigue detection in the experiment. To achieve early fatigue detection, we suggest capturing the real fatigue videos of tower crane operators during the operations instead of those with pretended facial expressions for the fatigue. (2) Due to the labelling segment level significantly affecting detection accuracy, we suggest that the human-verified labels be performed at the frame segment level. While labelling at every video frame takes much effort, labelling at every minute can be considered instead to save time and cost. Through the datasets with labels at the minute segment level, the trained model can also achieve relatively good accuracy. (3) Due to the variation of face poses, the facial videos captured side faces are less effective for fatigue detection than those captured the front faces. Given the high frequency of the head movement the crane operators would perform, installing both cameras to capture the front and side view respectively in the tower crane cabin is applicable for maximizing the detection quality. (4) For the complicated illuminations on the construction sites, we recommend installing the RGB and IR cameras for capturing the facial videos from different lighting spectrums regardless of daytime or night. It helps to establish robust and separated learning models to identify the different levels of fatigue anytime with varying illumination conditions.

Some limitations are identified in this study. Firstly, the comparative results are based on the simulation of crane operations. The realistic fatigue dataset for crane operators during actual operations could have other influential factors regarding quality that should be determined through further evaluations. Secondly, only a partial (but a large proportion though) of data among the three datasets is relabelled, due to the challenges of significant effort devoted to the relabelling process. The datasets should be relabelled completely at the frame level to achieve a unified evaluation criterion. Thirdly, the participants' characteristics, like age, years of driving experience, and gender, are not considered in this study due to the limited information provided from the available datasets. These can be taken into consideration in future work.

Acknowledgments

The authors would like to thank the Research Grants Council of Hong Kong for their Early Career Scheme funding support (PolyU 25221519). Partial research works, including the interviews and experiments participated by crane operators, are sponsored by the Start-up Research Project Scheme (No. P0001224) of the Hong Kong Polytechnic University. Also, thanks to You-Pin Construction Co., Ltd. for recruiting licensed crane operators to participate in the experiments of this research.

Reference

[1] T. Cheng and J. Teizer. Modeling tower crane operator visibility to minimize the risk of limited situational awareness. *Journal of Computing in Civil Engineering*, 28.3 (2014), 04014004, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000282.

[2] S. H. Han, S. Hasan, A. Bouferguène, M. Al-Hussein, and J. Kosa. Utilization of 3D visualization of mobile crane operations for modular construction on-site assembly. *Journal of Management in Engineering*, 31.5 (2015), 04014080, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000317.

[3] H.-L. Chi, Y.-C. Chen, S.-C. Kang, and S.-H. Hsieh. Development of user interface for tele-operated cranes. *Advanced Engineering Informatics*, 26.3 (2012), pp.641-652. https://doi.org/10.1016/j.aei.2012.05.001.

[4] R. L. Neitzel, N. S. Seixas, and K. K. Ren. A review of crane safety in the construction industry. *Applied Occupational and Environmental Hygiene*, 16.12 (2001), pp.1106-1117, https://doi.org/10.1080/10473220127411.

[5] V. W. Tam and I. W. Fung. Tower crane safety in the construction industry: A Hong Kong study. *Safety Science*, 49.2 (2011), pp.208-215, https://doi.org/10.1016/j.ssci.2010.08.001.

[6] X. Li, H.-L. Chi, W. Zhang, and G. Q. Shen. Monitoring and alerting of crane operator fatigue using hybrid deep neural networks in the prefabricated products assembly process. *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*. Banff, Canada, 2019, pp.680-687, https://doi.org/10.22260/ISARC2019/0091.

[7] A. Sahayadhas, K. Sundaraj, and M. Murugappan. Detecting driver drowsiness based on sensors: a review. *Sensors*, 12.12 (2012), pp.16937-16953, https://doi.org/10.3390/s121216937.

[8] M. Ngxande, J.-R. Tapamo, and M. Burke. Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques. *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*. IEEE, 2017, pp.156-161, https://doi.org/10.1109/RoboMech.2017.8261140.

[9] E. Tadesse, W. Sheng, and M. Liu. Driver drowsiness detection through HMM based dynamic modeling. *International Conference on Robotics and Automation (ICRA)*. 2014, pp.4003-4008, https://doi.org/10.1109/ICRA.2014.6907440.

[10] B. Reddy, Y.-H. Kim, S. Yun, C. Seo, and J. Jang. Real-time driver drowsiness detection for embedded system using model compression of deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp.121-128, https://doi.org/10.1109/CVPRW.2017.59.

[11] T. Åkerstedt and M. Gillberg. Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, 52.1-2 (1990), pp.29-37, https://doi.org/10.3109/00207459008994241.

[12] J.-M. Guo and H. Markoni. Driver drowsiness detection using hybrid convolutional neural network and long short-term memory. *Multimedia Tools and Applications*, 78.20 (2019), pp.29059-29087. https://doi.org/10.1007/s11042-018-6378-6.

[13] R. Ghoddoosian, M. Galib, and V. Athitsos. A Realistic Dataset and Baseline Temporal Model for Early Drowsiness Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp.178-187, https://doi.org/10.1109/CVPRW.2019.00027.

[14] W. Zhang, Y. L. Murphey, T. Wang, and Q. Xu. Driver yawning detection based on deep convolutional neural learning and robust nose tracking. *2015 International Joint

Conference on Neural Networks (IJCNN). IEEE, 2015, pp.1-8, https://doi.org/10.1109/IJCNN.2015.7280566.

[15] C.-H. Weng, Y.-H. Lai, and S.-H. Lai. Driver drowsiness detection via a hierarchical temporal deep belief network. *Asian Conference on Computer Vision*. 2016, pp.117-133, https://doi.org/10.1007/978-3-319-54526-4_9.

[16] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri. YawDD: A yawning detection dataset. *Proceedings of the 5th ACM Multimedia Systems Conference*. 2014, pp.24-28, https://doi.org/10.1145/2557642.2563678.

[17] I. J. Shin. Factors that affect safety of tower crane installation/dismantling in construction industry. *Safety Science*, 72 (2015), pp.379-390, https://doi.org/10.1016/j.ssci.2014.10.010.

[18] J. E. Beavers, J. Moore, R. Rinehart, and W. Schriver. Crane-related fatalities in the construction industry. *Journal of Construction Engineering and Management*, 132.9 (2006), pp.901-910, https://doi.org/10.1061/(asce)0733-9364(2006)132:9(901).

[19] A. Shapira and B. Lyachin. Identification and analysis of factors affecting safety on construction sites with tower cranes. *Journal of Construction Engineering and Management*, 135.1 (2009), pp.24-33, https://doi.org/10.1061/(ASCE)0733-9364(2009)135:1(24).

[20] G. Raviv, B. Fishbain, and A. Shapira. Analyzing risk factors in crane-related near-miss and accident reports. *Safety Science*, 91 (2017), pp.192-205, https://doi.org/10.1016/j.ssci.2016.08.022.

[21] G. Raviv, A. Shapira, and B. Fishbain. AHP-based analysis of the risk potential of safety incidents: Case study of cranes in the construction industry. *Safety Science*, 91 (2017), pp.298-309, https://doi.org/10.1016/j.ssci.2016.08.027.

[22] C. I. Council. (2010). *Guidelines on Safety of Tower Cranes*. Institution Hong Kong. Available from: http://www.cic.hk/files/page/50/Guidelines%20on%20Safety%20of%20Tower%20Cranes%20%28Version%202%29%20July%202010%20-%20e.pdf.

[23] G. Swaen, L. Van Amelsvoort, U. Bültmann, and I. Kant. Fatigue as a risk factor for being injured in an occupational accident: results from the Maastricht Cohort Study. *Occupational and Environmental Medicine*, 60.suppl 1(2003), pp.i88-i92, https://doi.org/10.1136/oem.60.suppl_1.i88.

[24] S. Park, F. Pan, S. Kang, and C. D. Yoo. Driver drowsiness detection system based on feature representation learning using various deep networks. *Asian Conference on Computer Vision*. 2016, pp.154-164, https://doi.org/10.1007/978-3-319-54526-4_12.

[25] I.-H. Choi, C.-H. Jeong, and Y.-G. Kim. Tracking a driver's face against extreme head poses and inference of drowsiness using a hidden Markov model. *Applied Sciences*, 6.5 (2016), pp.137, https://doi.org/10.3390/app6050137.

[26] K. Dwivedi, K. Biswaranjan, and A. Sethi. Drowsy driver detection using representation learning. *2014 IEEE International Advance Computing Conference (IACC)*. 2014, pp.995-999, https://doi.org/10.1109/IAdCC.2014.6779459.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 2012, pp.1097-1105, https://doi.org/10.1145/3065386.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014), https://arxiv.org/abs/1409.1556.

[29] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35.1 (2012), pp.221-231. https://doi.org/10.1109/TPAMI.2012.59.

[30] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp.1725-1732, https://doi.org/10.1109/CVPR.2014.223.

[31] F. Zhang, J. Su, L. Geng, and Z. Xiao. Driver fatigue detection based on eye state recognition. *2017 International Conference on Machine Vision and Information Technology (CMVIT)*. 2017, pp.105-110, https://doi.org/10.1109/CMVIT.2017.25.

[32] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9.8 (1997), pp.1735-1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[33] M. Akopyan and E. Khashba. Large-Scale YouTube-8M Video Understanding with Deep Neural Networks. *arXiv preprint arXiv:1706.04488* (2017), https://arxiv.org/abs/1706.04488.

[34] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib. Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication*, 71 (2019), pp.76-87, https://doi.org/10.1016/j.image.2018.09.003.

[35] T.-H. Shih and C.-T. Hsu. MSTN: Multistage spatial-temporal network for driver drowsiness detection. *Asian Conference on Computer Vision*. 2016, pp.146-153, https://doi.org/10.1007/978-3-319-54526-4_11.

[36] D. D. Chakladar, S. Dey, P. P. Roy, and D. P. Dogra. EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm. *Biomedical Signal Processing and Control*, 60 (2020), 101989, https://doi.org/10.1016/j.bspc.2020.101989.

[37] J. Lyu, Z. Yuan, and D. Chen. Long-term multi-granularity deep framework for driver drowsiness detection. *arXiv preprint arXiv:1801.02325* (2018), https://arxiv.org/abs/1801.02325.

[38] D. Babitha, M. Ismail, S. Chowdhury, R. Govindaraj, and K. Prakash. Automated road safety surveillance system using hybrid cnn-lstm approach. *International Journal of Advanced Trends in Computer Science and Engineering*, 9.2 (2020), pp.1767-1773, https://doi.org/10.30534/ijatcse/2020/132922020.

[39] M. Johns. The amplitude-velocity ratio of blinks: a new method for monitoring drowsiness. *Sleep*, 26.SUPPL.(2003). Available from: http://www.mwjohns.com/wp-content/uploads/2017/05/apss_2003_06_03_the_amplitude_velocity_ratio_of_blinks_a_new_method_of_monitoring_drowsiness_poster_69.pdf.

[40] L. K. McIntire, R. A. McKinley, C. Goodyear, and J. P. McIntire. Detection of vigilance performance using eye blinks. *Applied Ergonomics*, 45.2 (2014), pp.354-362, https://doi.org/10.1016/j.apergo.2013.04.020.

[41] M. Suzuki, N. Yamamoto, O. Yamamoto, T. Nakano, and S. Yamamoto. Measurement of driver's consciousness by image processing-a method for presuming driver's drowsiness by eye-blinks coping with individual differences. *2006 IEEE International Conference on Systems, Man and Cybernetics*. 2006, pp.2891-2896, https://doi.org/10.1109/ICSMC.2006.385313.

[42] H. Yin, Y. Su, Y. Liu, and D. Zhao. A driver fatigue detection method based on multi-sensor signals. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016, pp.1-7, https://doi.org/10.1109/WACV.2016.7477672.

[43] J. Jo, S. J. Lee, K. R. Park, I.-J. Kim, and J. Kim. Detecting driver drowsiness using feature-level fusion and user-specific classification. *Expert Systems with Applications*, 41.4 (2014), pp.1139-1152, https://doi.org/10.1016/j.eswa.2013.07.108.

[44] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23.10 (2016), pp.1499-1503, https://doi.org/10.1109/LSP.2016.2603342.

[45] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017), https://arxiv.org/abs/1704.04861.

[46] S. Arabi, A. Haghighat, and A. Sharma. A deep learning based solution for construction equipment detection: from development to deployment. *arXiv preprint arXiv:1904.09021* (2019), https://arxiv.org/abs/1904.09021.

[47] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28.10 (2016), pp.2222-2232, https://doi.org/10.1109/TNNLS.2016.2582924.

[48] K.-i. Funahashi and Y. Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6.6 (1993), pp.801-806, https://doi.org/10.1016/S0893-6080(05)80125-X.

[49] P. A. Gutiérrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28.1 (2015), pp.127-146, https://doi.org/10.1109/TKDE.2015.2457911.

[50] L. Gaudette and N. Japkowicz. Evaluation methods for ordinal classification. *Canadian Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2009, pp.207-210, https://doi.org/10.1007/978-3-642-01818-3_25.