Monitoring and Alerting of Crane Operator Fatigue Using Hybrid Deep Neural Networks in the Prefabricated Products Assembly Process

X. Li[a], H.L. Chi[a], W.F. Zhang[b], and Q.P. Geoffrey Shen[a]

[a]Department of Building and Real Estate, The Hong Kong Polytechnic University, China
[b]Ocean University of China, China
E-mail: shell.x.li@polyu.hk, hung-lin.chi@polyu.hk, zhangwenfeng@stu.ouc.edu.cn, geoffrey.shen@polyu.hk

Abstract –

Crane operators fatigue is one of the significant constraints should be monitored. Otherwise, it may lead to inefficient crane operations and safety issues. Recently, many deep neural networks have been developed for fatigue monitoring of vehicle drivers by processing the image or video data. However, the challenge is to distinguish the slight variations of facial features among still and motion frames (e.g., nodding and head tilt, yawning and talking). It can be exacerbated in the scenarios for crane operators due to their constant head moving to track the loads' position and recurrent communication (talking) with crane banksman. In contrast to previous approaches, which models spatial information and traditional temporal information for sequential processing, this study proposes a hybrid model can not only extract the spatial features by customized convolutional neural networks (CNN) but also enrich the modeling dynamic motions in the temporal dimension through the deep bidirectional long short-term memory (DB-LSTM). This hybrid model is trained and evaluated on the very popular dataset NTHU-DDD, and the results show that the proposed architecture achieves 93.6% overall accuracy and outperform the previous models in the literature.

Keywords –

Fatigue monitoring and alerting; Deep learning; Crane operator; Prefabricated construction

1 Introduction

In prefabricated construction, the prefabricated products become more and more complicated for assembly, with the evolution from components (light-weight, e.g., facade) and modules (large and heavy, e.g., volumetric precast bathroom) to pre-acceptance integrated units (larger and heavier, e.g., completed with finishes, fixtures, and fittings) [1]. Given this course of prefabricated products evolution, cranes, with their excellent transportation capacity, perform a decisive role in the assembly of prefabricated products by lifting them vertically and horizontally [2]. To achieve smooth crane operations, the crane operators should not only have enough physical strength but also be agile in the hearing, eyesight, and reflexes. As such, the operations and judgment of the crane operator will be a crucial factor for safety and productivity particular in the construction site of Hong Kong due to the high level of congestion and dynamics. However, the fatigue or drowsiness has been identified as the critical constraint in disturbing the operator's operations and judgment, which leads to the decreased attentiveness and vigilance, as well as casualties by collisions or falling loads [3,4]. In addition, Tam and Fung [3] revealed that around 60.5% of the crane operators would continue to work even feeling fatigue due to the long working hours (tight construction schedule) and about 52.6% of the crane operators are lack of breaks due to the inconvenient and narrow workspace (inconvenience of frequent in and out). Thus, automatically monitoring and warning the fatigue can provide timely support for crane operators, site superintendents and safety directors to make the scientific shifts and breaks.

Although there are seldom studies on developing fatigue monitoring and warning systems for the crane operator, numerous objective approaches have been proposed for detecting the fatigue or drowsiness of vehicle drivers from vehicle trajectory [5], physiological signal [6], and facial expression [7]. The first two approaches in crane operation can measure the fatigue by several parameters such as trolley movement speed, loads path deviation, jib rotation speed, heart rate, electroencephalogram (EEG), electrooculogram (EOG), electromyogram(EMG), and electrocardiogram (ECG). These two methods have shown a good accuracy when monitoring physical fatigue of vehicle drivers. However, crane operation trajectory may be affected by other factors (e.g., operation errors due to inexperience, inefficient communication with site signaller) and the physiological signal should be collected by an annoying

and invasive way to crane operators. Thus, monitoring the fatigue reflected by facial expressions (e.g., eye state, yawning, nodding) can be a more convenient, fast-speed and cost-effective approach. This kind of approach can analyze the facial features extracted from the videos/images of crane operators, and it performs a high accuracy after the boosting of various deep neural networks as it facilitates the computer to learn by itself for capturing the key features. For example, Zhang et al. [8] adopted the convolutional neural network (CNN) to detect the yawning by using the features in nose region instead of mouth area due to the head turnings of vehicle drivers. However, it is still difficult to distinguish easy-to-confuse fatigue states, such as blinking and closing eyes. Huynh et al. [9] provided a more practical solution with the 3D-CNN by considering the broader features on the face and temporal information (sequence of video frames). However, it is still a challenge to distinguish the fatigue states with long-term dependencies, such as yawning and talking. Guo and Markoni [10], and Lyu et al. [11] improved the learning model on the temporal information by integrating CNN with Long Short-Term Memory (LSTM) network, which is a type of recurrent neural network (RNN) that can distinguish the states with long-term dynamical features over sequential frames. However, the potential of CNN-LSTM is far from being fully exploited in the domain of driver/operator fatigue monitoring. The primary limitation in previous studies on CNN-LSTM in fatigue monitoring is that the long-term dependencies of periodic fatigue behavior (e.g., distinguish nodding and head tilt along with loads movements, yawning and talking) are learned from positive-sequence video frames considering only forward dependencies, while backward dependencies learned from reverse-order frames has never been explored that means some useful information may be missed.

To address this issue for improving the accuracy in monitoring and alerting of crane operator fatigue, this study develops a hybrid deep neural network by integrating CNN with deep bidirectional LSTM (DBLSTM) network. The specific objectives of this study are: (1) to accurately detect and align the facial regions with critical fatigue features; (2) to extract the effective facial fatigue features on single-frame images; (3) to distinguish the fatigue state by mining bidirectional temporal clues of sequential features.

2 Literature Review

Crane operator executes the repetitive lift tasks under the fatigue state in a complex construction environment may lead to catastrophic casualties as same as the vehicle drivers. There are apparent signs that suggest 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 [12]. As the facial features of operator/driver in a fatigued state are significantly different from that of the conscious state, the real-time monitoring the operator/driver's face by the camera can be an efficient, non-invasive and practical approach to alert the drowsiness and avoid the accidents [13]. PERCLOS (percentage of eyelid closure over the pupil over time) is a reliable measure to monitor the fatigue [14]. In addition, numerous machine learning-based approaches have also been applied to fatigue monitoring. For example, Mbouna et al. [15] developed an approach to extract the visual features from the eyes and head pose of the drivers, and then support vector machines (SVMs) was used to classify the fatigue levels. Choi et al. [16] trained the hidden Markov models (HMMs) to model the temporal behaviors of head pose and eye-blinking for identifying whether the driver is drowsy or not. However, these approaches relied on hand-crafted features which have shown limited efficacy in real-time monitoring and can be inaccurate when driver/operator wear the sunglasses or under considerable variation of illumination conditions [17]. Concurrently, features learned from unlabelled data based on the deep neural networks such as the convolutional neural network (CNN) have been proved to have a significant advantage over hand-crafted features in real-time monitoring of fatigue [14].

CNN is the class of deep and feed-forward neural networks that involves three main elements including local receptive fields, shared weights, and spatial or temporal pooling [18]. The process of fatigue monitoring and alerting by CNN is the same as other machine learning-based methods that can be shown in Figure1. The previous studies regarding fatigue monitoring and alerting by using CNN related models have also been summarized in Table 1. CNN was first applied to fatigue monitoring as the features extractor of static facial fatigue images by Dwivedi et al. [19]. Then, Zhang et al. [8] used the CNNs 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.[20] 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 the fatigue states are dynamic (e.g., yawning, nodding) and it is difficult to distinguish whether the driver/operator is yawning or talking when only capturing their open mouths, a 3D CNN was proposed to capture the motion information of numerous adjoining frames from videos, and 3D filters

(kernel) were adopted to extract spatiotemporal features [9]. Furthermore, Part et al. [17] integrated three existing CNN-based models including AlexNet, VGG-FaceNet, and FlowImageNet in terms of their efficiency in the extraction of image features, facial features, and temporal features. However, these methods can only extract features with fixed temporal length, and the 3D convolution processes may spend numerous resources and time to impede the real-time monitoring.

Long Short-Term Memory networks (LSTMs) has been proved to be effective in learning long-term temporal dependencies by solving the exploding and vanishing gradient problems that is a Gordian knot for the traditional recurrent neural network (RNN) [21]. And an LSTM comprises typically 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. Thus, the integration of CNN and LSTM can be an alternative in fatigue monitoring and alerting. 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. And several refinement techniques help them achieve the high accuracy such as reducing the hidden layer of LSTM [10], noisy smoothing in post-processing [22], and alignment technology to learn the most critical fatigue information [11]. However, to improve the accuracy, all information included in time series data should be entirely employed. The frames of video are sequentially fed into an LSTM that lead to an information flow with positive direction from time step t-1 to t along the chain-like structure. Therefore, the LSTM can only utilize the forward dependencies, and it is very likely that valuable information is filtered out or not efficiently passed through the chain-like gated structure [23]. Thus, it may enrich the temporal features by considering the backward dependencies. Moreover, the facial expressions of fatigue can be periodical and regular, and even short-term periodicity such as nodding can be detected. Learning the periodicity of time series data, particularly for recurring fatigue patterns, from both forward and backward temporal information can improve the performance of fatigue monitoring and alerting. However, to the authors' knowledge, few studies on crane operator fatigue monitoring considered the backward dependencies. To fill this gap, a deep bidirectional LSTM (DB-LSTM) is integrated into the CNN to form the architecture of fatigue monitoring and alerting system.

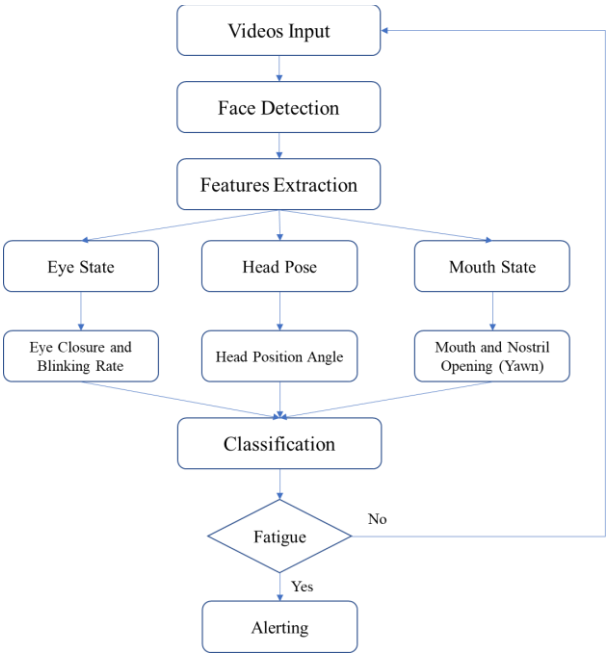

Figure 1. The machine learning-based process of facial fatigue monitoring and alerting

Table 1. The summary of studies by using deep neural networks for fatigue monitoring

Research	Techniques	Database	Accuracy
Dwivedi et al.2014	CNN, Viola and Jones algorithm	Customized	78%
Zhang et al. 2015	CNN, Kalman filter with track-learning-detection (TLD)	YawDD	92%
Huynh et al. 2016	3D CNN, Gradient Boosting	NTHU	87.46%
Park et al. 2016	AlexNet, VGG-FaceNet, FlowImageNet	NTHU	73.06%
Shih and Hsu, 2016	VGG-16, LSTM	NTHU	85.52%
Reddy et al. 2017	Multi-Task Cascaded CNN,	Customized	89.50%
Guo&Markoni, 2018	MTCNN,VGG-11,LSTM	NTHU	84.85%
Lyu et al. 2018	Multi-granularity CNN, LSTM	NTHU	90.05%

3 Proposed Solution

Figure 2 shows the architecture of the proposed hybrid deep neural networks, which comprises three steps and each step maps to a specific model. Firstly, the multi-task cascaded convolutional networks (MTCNN) are adopted as the face detector to locate and align the facial area in each frame of the video. Secondly, the customized CNN model is designed to extract facial

fatigue features from individual-frame images. Finally, a sequence of features within a specific time interval is fed into DB-LSTM to model the temporal variation of fatigue. And the Gaussian smoothing is adopted to reduce the noise and improve the fatigue monitoring performance. Each step of the proposed method is detailed in the following sections.

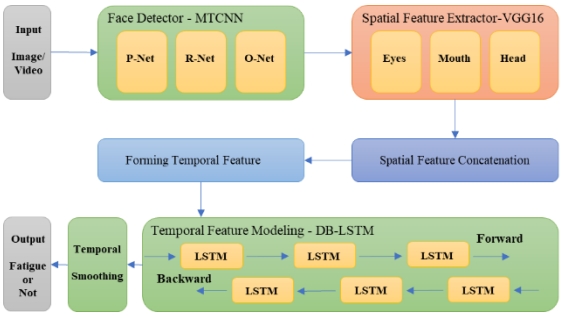

Figure 2. The architecture of the hybrid deep neural networks for fatigue monitoring and alerting

3.1 Face Detection

Precisely detecting and aligning the facial area of crane operator from an image is very critical to achieve efficient extraction of facial fatigue features and fatigue recognition. One of the famous face detectors proposed by Viola and Jones, [24] uses Haar-Like features and AdaBoost to train cascading classifiers, which achieves high detection rate in a real-time manner. However, previous studies have proved and indicated that the accuracy and efficiency of this face detector might reduce with large variations of facial regions [9,16]. These challenges could be exacerbated for face detection and alignment during crane operations in the real-world situations, such as the large pose variations of the operator who should change pose along with the moving loads, extreme lightings or darkness in operation cabin, and occlusions in front of the face. To fill this gap, the multi-task cascaded convolutional networks (MTCNN) proposed by Zhang et al., [25] shows the significant performance improvement in both accuracy and efficiency compared with other face detectors. This study adopts MTCNN to conduct the face detection and face alignment tasks with several stages. Firstly, the input images with various scales should be resized to build an image pyramid. Secondly, a shallow CNN (P-Net) with the input size of 12×12 to fast generate the candidate facial windows that are calibrated based on the bounding box regression vectors, and the highly overlapped candidates are fused by using non-maximum suppression (NMS). Thirdly, a complex CNN (R-Net) with the input size of 24×24 is adopted to reject the non-facial candidates with the same process of calibration and fusion. Finally, a more powerful CNN (O-Net) with the input size of 48×48 is applied to refine the results and produce five landmark points including positions of left-eye, right-eye, nose, left-lid-end, and right-lip-end.

3.2 Spatial Features Extraction

The objective of the features extraction is to learn a CNN-based spatial-domain feature extraction model E for capturing fatigue features F from the individual facial images I. As the feature extraction model E would go through each individual image in I, the extracted F should be general and robust to different input noises. Thus, this study chooses VGG-16 as our basic model which has achieved good performance in various datasets of image recognition [26]. On the basis of original VGG-16, several improvements are conducted to balance the efficiency and accuracy for extracting fatigue feature in a real-time condition. Figure 3 demonstrates the improved VGG-16 architecture V. The original VGG-16 which includes 13 convolutional layers (grouped into Conv 1-5), 5 max-pooling layers (pool 1-5), and 3 fully connected feedforward network layers. However, the input of this study has a smaller size image (64×64 RGB images) than the original VGG-16 (224×224 RGB images), which means the number of parameters can be reduced by using smaller fully connected (Fc) network layers (Fc-6, Fc-7, binary classifier) to avoid over-fitting in the improved VGG-16. Given the input to the improved VGG-16 is a fixed-size 64×64×3 face image, the features both in max pooling 5 and max-pooling 3-4 can be used to obtain the discriminative representation. This considers the fact that forward layers of CNN include more detailed information, while the backward layers summarize the global information. This improvement can be beneficial for improving the accuracy of extracting the small region features that are easily ignored by max-pooling, such as the eyes. To this end, a 1×1 convolutional layer is applied into each of pool 3-5 to approximate the Fc 6 by generating three vectors with the same depth (e.g., 256 in this study). This approximation strategy can not only reduce the number of parameters of Fc layers but also facilitate the Fc layers to extract fatigue-related features by pooling operation automatically. The pooled vectors are concatenated to feed into Fc 7 with fewer parameters to extract the more critical features, which forms the F.

To enable the faster and stable training process in generating the feature extraction model E with good generalization, another improvement for VGG-16 is to use Batch Normalization (BN) [27]. BN is a kind of feature scaling technique that can normalize the sample mean and variance of hidden units before or after the process of activation functions over mini-batch data.

This normalization process helps lessen the internal covariate shift for allowing using the larger learning rates. Meanwhile, the mini-batch including various samples may lead to randomness, which can reduce the risk of over-fitting. In this study, there are 5 BN layers placed before max-pooling layer and Fc layer. Lastly, a binary classifier is placed after Fc7 to predict the fatigue score Y. Given both Y and ground truth label L ϵ {0,1}, the cross-entropy loss function with the Adam optimizer is adopted to minimize the loss. If the Y is larger than 0 and is close to 1, the fatigue degree of the input is higher, and vice versa.

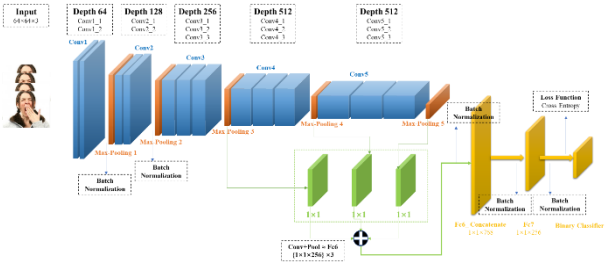

Figure 3. The architecture of CNN-based spatial-domain feature extraction model

3.3 Temporal Features Extraction

Although the feature extractor E has already enabled to predict the fatigue score of each 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 can be meaningful to consider both backward and forward information in the sequential frames. To this end, the deep bidirectional long short-term memory (DB-LSTM) is applied to model the temporal features F. DB-LSTM can process the sequential data from two directions by two separate hidden layers and then feed them into the same output layer. The outputs of forward and backward layers (as shown in Figure 4(b)) are both computed by using the basic structure of standard LSTM, See Figure 4.

DB-LSTM has a memory cell to save the state vector which is the sequence of the past or future input data. The current state can be updated on the basis of the current input, output, and the previous state saved in that "cell." DB-LSTM has a gated structure which allows the network to forget the previous state saved in cells or to update the latest state based on the new input data. At time t, the input gate vector, forget gate vector, output gate vector and the state of the memory cell can be denoted as it, ft, ot, and ct respectively. then ct can be updated by the equation (1)-(6).

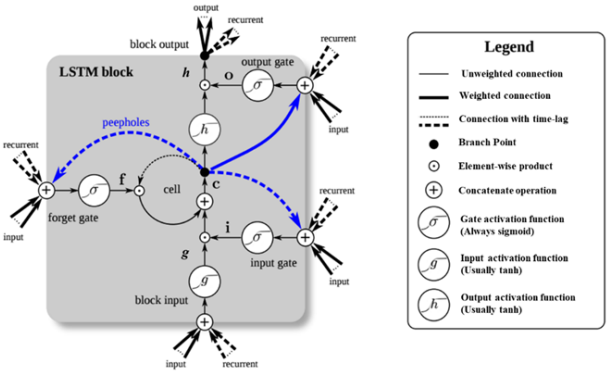

Figure 4 (a). The structure of the standard LSTM [28]

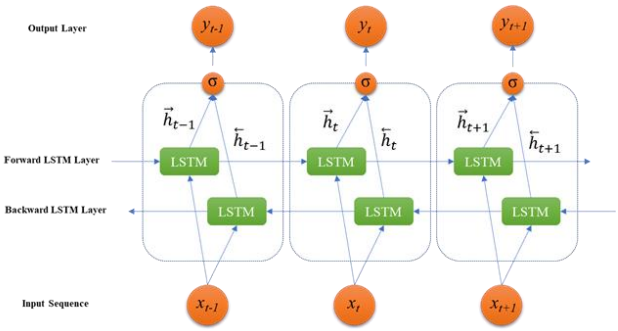

Figure 4 (b). The architecture of the DB-LSTM

$$i_t = \sigma_i(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma_f(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma_o(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (3)$$

$$g_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$

$$h_t = o_t \odot tanh(c_t) \quad (6)$$

Where x_t is the input and σ is the gate activation function, which usually is the sigmoid function. g_t is the state update vector that has activation function ''tanh'' (hyperbolic tangent function) and is computed from the input of the current state and previous state. Forget gate f_t allows the LSTM to forget its previous memory cell c_{t-1} or further memory cell c_{t+1}, and the output gate o_t adopts a transformation to the current memory cell to produce the hidden state h_t. For three gates, the gate can accept the input vector only if the gate value is 1 and reject the input vector when the gate value is 0. Weight matrices $\boldsymbol{W}$ and biases $\boldsymbol{b}$ are the trained parameters. $\odot$ indicates the element-wise product with the gate value. Then, the vector $\boldsymbol{Y_t}$ in feature sequence $\boldsymbol{F}$ is the

concatenated vector by combining the outputs of forward and backward processes as follows:

$$Y_t = \vec{h}_t \oplus \overleftarrow{h}_t \qquad (7)$$

Where $\oplus$ represents the concatenate operation.

In this study, each video can be randomly sampled as the training data by dividing it into numerous video clips with fixed length 50. The DB-LSTM temporal network includes 64 hidden units to predict the refined fatigue score $\boldsymbol{Y_t}$ of each frame (t=1,…,50). The cross-entropy loss function with the Adam optimizer is still applied to minimize the loss.

In the previous stages, both spatial network (CNN) and temporal network (DB-LSTM) are applied to predict the fatigue score of each frame. However, there are still certain noises during the testing on the validation set. In order to achieve a better performance of accuracy, the post-processing techniques including Gaussian smoothing, moving mean/median filtering can be adopted to "smoothing" the predicted fatigue scores.

4 Results and Conclusions

Figure 5 demonstrates the average loss among 20 videos of the training set (orange line), and the evaluation set blue line). The spatial features extraction model E already achieves 85.82% accuracy of fatigue even though its prediction is merely based on a single frame. Figure 6 represents the comparison of DB-LSTM and LSTM on accuracies and convergent performance in the evaluation set. The temporal network LSTM models the temporal variation of the fatigue status, and thus improves the accuracy of fatigue to 92.20%. It is worth noting that a longer clip length T during testing achieves higher accuracies. Finally, after adopting DB-LSTM, it achieves 93.60 % accuracy.

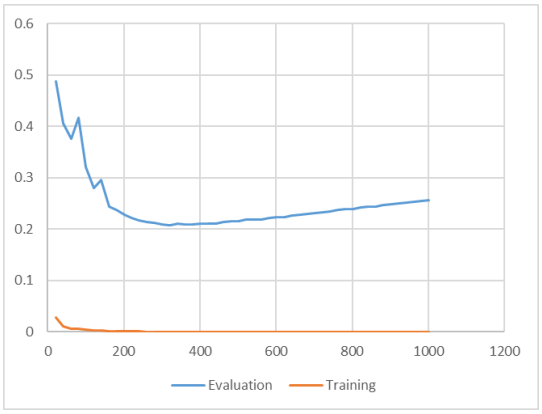

Figure 5. Loss curve of DB-LSTM for both training set and evaluation set

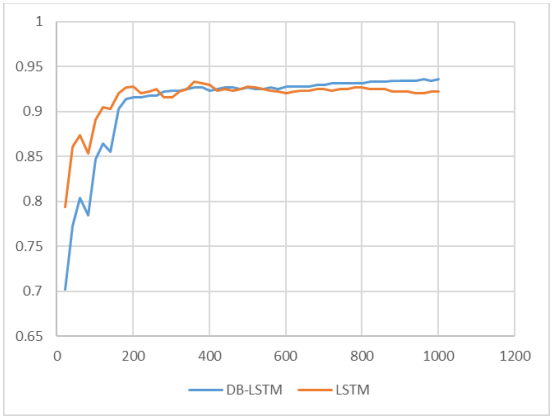

Figure 6. The comparison of DB-LSTM and LSTM on accuracies and convergent performance in the evaluation set

Table 3 represents the average F1 scores and accuracies on the evaluation set. In term of accuracy, the proposed hybrid neural network works pretty well under the sunglasses scenario (98.82%) and Non-Glasses scenario (95.41%). In terms of F1 score, the balanced F1 score among all scenarios shows that the proposed method does not make a biased prediction.

Table 3. Average F1 scores and accuracies for different scenarios

	F1-Score	Accuracy	Number of clips
Night_nonglasses	0.8080	0.8800	125
Night_glasses	0.6061	0.8839	112
Glasses	0.9617	0.9440	125
Non_glasses	0.9655	0.9541	109
Sunglasses	0.9870	0.9882	170
All	0.9286	0.9360	641

This study proposes a hybrid neural network to monitor and alert the fatigue status of the crane operator on the videos. The improvements and contributions in this study are threefold: (1) expand the vehicle driver drowsiness detection to the crane operator fatigue monitoring and alerting; (2) to detect and align the face, and extract the spatial features, the customized CNN is developed based on the baseline models which have excellent performance; (3) a deep bidirectional LSTM (DB-LSTM) is developed by considering both forward and backward dependencies to model the temporal pattern, which can learn compositional representations in space and time. The experiment results indicate that the effectiveness of the proposed hybrid neural network in comparison with several state-of-the-art methods. Further improvements and extensions can be made based on this study. Firstly, the dataset for crane operators should be established instead of using datasets

from the vehicle drivers. Additionally, devising more powerful features by combining multiple signals such as ECG, human audio, other physiological signals can be considered to achieve better accuracy and efficiency in fatigue monitoring.

Acknowledgments

This research was supported by National Key R&D Program of China (No.2016YFC070200504).

References

[1] Han, S. H., Hasan, S., Bouferguène, A., Al-Hussein, M., & Kosa,J. Utilization of 3D visualization of mobile crane operations for modular construction on-site assembly. *Journal of Management in Engineering*, *31*(5), 04014, 2014.

[2] Chi, H. L., Chen, Y. C., Kang, S. C., & Hsieh, S. H. Development of user interface for teleoperated cranes. *Advanced Engineering Informatics*, *26*(3), 641-652, 2012.

[3] Tam, V. W., & Fung, I. W. Tower crane safety in the construction industry: A Hong Kong study. *Safety Science*, *49*(2), 208-215, 2011.

[4] Marquez, A., Venturino, P., & Otegui, J. Common root causes in recent failures of cranes. *Engineering Failure Analysis*, 39, 55-64. 080,2014.

[5] Thiffault, P., & Bergeron, J. Monotony of road environment and driver fatigue: a simulator study. *Accident Analysis & Prevention*, *35*(3), 381-391, 2003.

[6] Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue, and drowsiness. *Neuroscience & Biobehavioral Reviews*, *44*, 58-75, 2014.

[7] Ji, Q., & Yang, X. Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. *Real-time Imaging*, *8*(5), 357-377, 2002.

[8] Zhang, W., Murphey, Y. L., Wang, T., & Xu, Q. (2015, July). Driver yawning detection based on deep convolutional neural learning and robust nose tracking. In *Neural Networks (IJCNN), 2015 International Joint Conference on* (pp. 1-8). IEEE.

[9] Huynh, X. P., Park, S. M., & Kim, Y. G. Detection of driver drowsiness using the 3D deep neural network and semi-supervised gradient boosting machine. In *Asian Conference on Computer Vision* (pp.134-145). Springer, Cham, 2016.

[10] Guo, J. M., & Markoni, H. Driver drowsiness detection using the hybrid convolutional neural network and long short-term memory. *Multimedia Tools and Applications*, 1-29, 2018.

[11] Lyu, J., Yuan, Z., & Chen, D. Long-term multi-granularity deep framework for driver drowsiness detection. arXiv preprint arXiv:1801.02325, 2018.

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

[13] Shi, S. Y., Tang, W. Z., & Wang, Y. Y. A Review on Fatigue Driving Detection. In *ITM Web of Conferences* (Vol. 12, p. 01019). EDP Sciences, 2017.

[14] Zhang, F., Su, J., Geng, L., & Xiao, Z. Driver fatigue detection based on eye state recognition. In *Machine Vision and Information Technology (CMVIT), International Conference on* (pp. 105-110). IEEE, 2017.

[15] Mbouna, R. O., Kong, S. G., & Chun, M. G. Visual analysis of eye state and head pose for driver alertness monitoring. IEEE transactions on intelligent transportation systems, 14(3), 1462-1469, 2013.

[16] Choi, I. H., Jeong, C. H., & Kim, Y. G. Tracking a driver's face against extreme head poses and inference of drowsiness using a hidden Markov model. *Applied Sciences*, *6*(5), 137, 2016.

[17] Park, S., Pan, F., Kang, S., & Yoo, C. D. Driver drowsiness detection system based on feature representation learning using various deep networks. In *Asian Conference on Computer Vision* (pp. 154-164). Springer, Cham, 2016.

[18] LeCun, Y., & Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

[19] Dwivedi, K., Biswaranjan, K., & Sethi, A. Drowsy driver detection using representation learning. In *2014 IEEE International Advance Computing Conference (IACC)*(pp. 995-999).

[20] Reddy, B., Kim, Y. H., Yun, S., Seo, C., & Jang, J. Real-time Driver Drowsiness Detection for Embedded System Using Model Compression of Deep Neural Networks. *Comput. Vis. Pattern Recognit. Work*, 2017.

[21] Hochreiter, S., & Schmidhuber, J. Long short-term memory. *Neural Computation*, *9*(8), 1735-1780, 1997.

[22] Shih, T. H., & Hsu, C. T. MSTN: a multistage spatial-temporal network for driver drowsiness detection. In *Asian Conference on Computer Vision* (pp. 146-153). Springer, Cham, 2016.

[23] Cui, Z., Ke, R., & Wang, Y. Deep Bidirectional and Unidirectional LSTM Recurrent Neural

Network for Network-wide Traffic Speed Prediction. *arXiv preprint arXiv:1801.02143*, 2018.
[24] Viola, P., & Jones, M. Robust real-time face detection. In *null* (p. 747). IEEE, 2001.
[25] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499-1503, 2016.
[26] Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[27] Ioffe, S., & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
[28] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222-2232, 2017.