Proceedings of the 34th European Modeling & Simulation Symposium (EMSS), 029
19th International Multidisciplinary Modeling & Simulation Multiconference

ISSN 2724-0029 © 2022 The Authors.
doi: 10.46354/i3m.2022.emss.029

Mobile crane signalman static hand signals classification framework using deep convolution neural network

Asif Mansoor[1,*], Shuai Liu[1], Ghulam Muhammad Ali[1], Ahmed Bouferguene[2], Mohamed Al-Hussein[1], and Soda[3]

[1]Department of Civil and Environmental Engineering, University of Alberta, 9105 116 street, Edmonton, T6H 2W2, Canada
[2]Campus Saint-Jean, University of Alberta, Edmonton, T6C 4G9, Canada
[3]Department of Computer Science, University of Turbat, Pakistan

[*]Corresponding author. Email address: amansoor@ualberta.ca

Abstract

Cranes are the need of every construction site as the construction paradigm is shifting from traditional (on-site) methods toward an off-site (modularization) approach. The communication between the crane operator and the crane signalman plays a significant role to complete the construction project safely and efficiently. The communication between crane operator and signalman relies on hand signals and two-way radio communication systems. However, these means of communication are not reliable in modern construction as the construction sites are more congested and noisy. Any miscommunication may lead to a disastrous accident on the construction site. The recent advancement in information technology can assist to add more layers of communication in the crane Industry. This paper presents a framework that uses deep convolutional neural network (DCNN) architecture for static hand signal classification using the crane signalman hand signals dataset. The DCNN model was developed to classify 18 different crane signalman hand signals. The model was trained, validated, and tested using a dataset of 8133 images, and achieved average accuracies of 89.1% and 84.6% for the training dataset and the validation dataset, respectively. The precision, recall, and F1 score in the test dataset were recorded as 81.5%, 81.8%, and 81.7%, respectively. The model is further validated with real-time hand signals classification and an accuracy of 87.9% is achieved. This developed framework can be used as another layer of communication with the current state of practice to reduce the communication error between crane signalman and operator.

Keywords: Crane signalman hand signals; deep convolutional neural network, communication, classification

1. Introduction

The communication between the crane operator and the signalman relies on hand signals and two-way radio communication systems. The crane industry has been using universal hand signals for decades to give direction to the crane operator for safe crane operation (Everett and Slocum 1993). Hand signals are the fastest and most reliable way to communicate a message when the crane operator has a direct line of sight with the signalman and the operator then operates the crane based on the direction given by the signalman. However, when the signalman is far away

from the crane operator, the crane operator will not be able to clearly distinguish the signalman's hand signals; while at other times the crane operator's line of sight will be obstructed due to construction site congestion. These limitations make this method ineffective and unsafe (Everett and Slocum 1993; Shapira et al. 2008). To overcome the limitations, the crane industry has used two signalmen at the same time such that the second signalman copies the main signalman's signals and transmits these to the operator; however, this process is less productive and cannot be 100% reliable due to the potential for the signalman to misinterpret the signals or miscommunicate them to the crane operator, which may lead to a disastrous accident(Fang and Cho 2016). Another means of communication in the crane industry is the use of a two-way radio communication system. Typically, this two-way radio communication system is used on the site when the crane operator's direct line of sight/vision is blocked by an obstacle or when the signalman is far away from the operator and he cannot see the signalman clearly; however, a two-way radio communication system needs to be maintained on a dedicated channel that is available at all times (Zekavat et al. 2014). This system can be viewed as an alternative to the hand signaling system under certain circumstances and, potentially, as an extra layer of safety. A two-way radio communication system cannot be 100% reliable when the construction site is noisy, such as when there is drilling on-site and the operator cannot hear the signalman clearly (Zavichi and Behzadan 2011; Mansoor et al. 2020). Secondly, while communicating using two-way radios, one hand needs to be used to push the talk button to send the voice message while the other hand must be used for signaling, which may cause a miscommunication error and can be dangerous (Zekavat et al. 2014). Another drawback of the two-way radio system is when a problem occurs with the dedicated channel being used by the operator and signalman because the whole crane operation is halted until the problem is resolved, which causes delays, productivity losses, and safety risks (Zavichi and Behzadan 2011).

To overcome the aforementioned limitations, information technology can provide another layer of safety by making the communication between signalman and crane operator more efficient and more accurate by using deep convolutional neural network (DCNN). The objectives of the framework are to develop a DCNN model, to train the model with an image dataset of crane signalman hand signals, to achieve a high level of accuracy in classifying images of crane signalman hand signals in training, validation, and test datasets, and to further validate the model for real-time crane signalman hand signals classification. The proposed framework is capable of classifying the crane signalman hand signals in real-time. The benefit of using this framework is that it assists communication between the crane operator and the signalman.

The present study is organized as follows: Section 2 Introduces related work in context of improving the communication between crane operator and signalman. Section 3 describes the methodology and architecture of the proposed deep convolutional neural network model. Section 4 and 5 presents the model optimization techniques and data preparations. Section 6 evaluates the Implementation and model performance and describe the model real-time

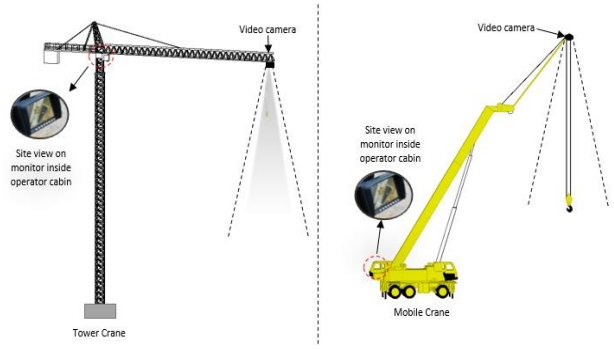

Figure 1. Camera-based vision system for tower and mobile cranes

classification results of the proposed model. Section 7 contain the conclusion, limitations of the present study and related future work.

2. Related work

To improve safety and communication in the context of crane operation, researchers have done a significant amount of work. Camera-based vision systems have been developed that enable the operator to monitor the construction site while operating the crane. Researchers have also developed sensors to detect hazards and dangerous situations on the construction site. These technologies are quite beneficial in terms of improving safety and communication on construction sites, but due to the limitations of these technologies, they are not yet widely employed in the crane industry. These technologies are discussed below in detail.

2.1. Camera-based systems

In crane industry, in particular, can benefit from improved communication and safety systems; therefore, researchers have developed camera-based systems as shown in Figure 1. (Shapira et al. 2008) developed a video monitoring system that not only improves safety in the construction site but also increases productivity by 11 to 26%. The noted results were based on 2400 time complete delivery of payload from picking to dropping the load to their allotted locations. The system consists of a video camera mounted to the top of the tower crane to show a live video feed of the site that focuses on the signalman (Shapira et al. 2008). A monitor in the crane cabin allows the operator to see the live video of the signalman as well as the site. According to (Rosenfeld 1995), video cameras attached to the cranes can be

useful both in terms of efficiency and safety improvements as the live site vision enables the operator to make judgments on-site without any hesitation.

The drawback of this technology is that the camera shows only 2-dimensional images without any perception of depth. The operator has difficulty accurately determining the distance of the load to the ground, which can lead to serious accidents on the site(Shapira et al. 2008). CRANIUM is another camera-based technology, developed by Everett (Everett and Slocum 1993), that has been found to improve both safety and cost-effectiveness because it eliminates the need for the second signalman (who is responsible for copying the hand signals of the first signalman at the lifting point and communicating these to the crane operator) (Everett and Slocum 1993). In this system, a camera is fixed to the top of the boom and a monitor is placed in the operator's cabin. From the monitor, the crane operator can see the loading area as well as the signalman (Everett and Slocum 1993). Stoneridge-Orlaco and HoistCam are camera and monitor manufacturers providing camera-based solutions for live video feed for tower, telescopic and crawler cranes to improve the communication, safety, and efficiency of crane operation on construction sites. The disadvantage of these camera-based systems is that sometimes while the crane is in operation, a part of the crane, such as a hook or sling, moves in front of the camera which obstructs the operator's view and the operator is unable to see the signalman or the target area. This drawback can slow down the operation and increase the safety risks for the workers on the site (Everett and Slocum 1993).

2.2. Sensor-based systems

Another means to accomplish improved safety and communication in the crane industry is the application of sensor-based systems. (Li et al. 2013) introduced RFIDs to track construction workers on-site with the help of a GPS tracking system. This system is used to limit the movement of the workers because only authorized workers were allowed in the danger zone during any crane lift operation. By automatically detecting unauthorized workers, the system sends an alert to warn authorized management personnel. This system was approved by the management and staff on the site (Li et al. 2013). (Fang and Cho 2016) developed a sensor-based framework that was capable of providing real-time safety assistance for mobile crane lift operations on a construction site. The sensors were attached to a 70-ton telescopic crane and were responsible for detecting nearby objects and warning the crane operator of any dangers. This framework was implemented on the site and the responses from the working crew were recorded. Most of the crew was satisfied with the developed framework and responded that the framework can be helpful during blind lifts by responding early before the crane comes into contact with any object (Fang and Cho 2016). To improve the two-way radio communication system between crane operator and signalman, (Zekavat et al. 2014) developed a camera-based vision system with a wireless microphone to monitor the blind lifts. In this system, a laptop was placed above the eye level of the operator in the crane operator cabin, where the operator have to look up the laptop to make decision while moving the load. This system improved the communication and visibility of the site during crane operation (Zekavat et al. 2014). On the other hand, the limitations of sensor-based technologies are the accuracy of the system, and the measurement error and setup error of the system makes these systems less trustworthy because a small margin of error can cause serious accidents on the construction site (Everett and Slocum 1993). With respect to implementing RFIDs for the purpose of worker tracking, the workers responded that they have privacy concerns, in particular, that their productivity will be judged based on their movements; this system was noted to have accuracy issues as well (Li et al. 2013).

3. Methodology and CNN architecture

In recent years, rapid improvements in computation power have led to the development of deep learning algorithms for image classification, notably, convolution neural networks (CNNs). A convolution neural network extracts the features from the input image pixels and classifies the images with high accuracy and generalization capabilities. In the civil engineering domain, convolution neural networks have been used for the classification and detection of equipment on construction sites, and construction workers wearing hardhat (Arabi et al. 2019; Gu et al. 2019; Hu et al. 2019; Wang et al. 2020; Wu et al. 2019), of defects in sewer pipes (Yin et al. 2020), and of defects in concrete surfaces (Cha et al. 2017), for example.

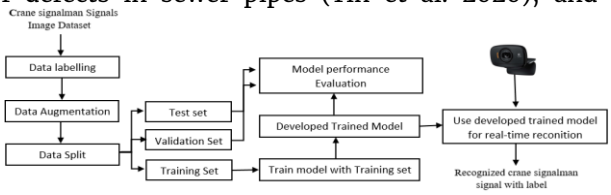

Figure 2. Overview of the deep learning based framework

In the present study, a deep convolution neural network approach is developed to classify crane signalman static hand signals in real-time. The developed DCNN is trained, validated, and tested using a dataset of 8133 augmented and non-augmented images that are collected by individuals from the research team. An overview of the developed deep learning-based framework is given in Figure 2. The crane signalman hand signals dataset containing all 18 hand signals necessary for crane operation serves as the original data source for the signalman hand signals. The crane signalman hand signals are based

on the occupational health and safety (OHS) crane hand signals (Occupational Safety and Health Administration., 2009), as shown in Figure 3.

In order to improve communication, the crane signalman hand signals must be classified in real-time. To accomplish this, a deep learning approach is developed and is consists of two main parts. First data is collected and pre-processed and the second is the development of the deep convolutional neural network model that is discussed in detail in section 3.1 and 3.2

3.1. Data collection and pre-processing

Since there is no accessible dataset available for crane signalman hand signals, the dataset needed for this work is collected by taking 6507 images of individuals from the research team doing all 18 crane signalman hand signals in all possible angles (0°to 360°), and positions (right, left, front and back sides of signalman). The camera is set up at multiple distances (5meter to 20meter) to record the image dataset in different environments (sunny and cloudy). While creating the dataset, data balancing is taken into account, which means the portion of images in the dataset showing each one of the 18 hand signals should be 5.5% of the whole dataset. However, in the collected dataset, 7.51% of images depict the emergency stop hand signal, which is the maximum portion, and the minimum portion of images, 3.64%, depict the travel hand signal as shown in Table 1. Therefore, the amount of data provided to train the model is sufficient for the model to learn the features from all hand signals in the dataset. The number of samples collected for each hand signal can be seen in Table 1.

All images have a resolution of 1280×720 pixels. The deep convolution neural network can be trained using images of any resolution; however, a higher resolution means more features will be extracted from images, which increases the computational complexity and the processing time. To reduce the processing time and computational load, all images are scaled down to a 280×280 pixel resolution for further processing.

Table 1. Image samples collected for each hand signal

Standard Hand signals	Number of sample images	Percentage (%)
Hoist	486	7.47
Lower	411	6.32
Use main hoist	231	3.55
Use whipline	333	5.12
Boom up	477	7.33
Boom down	453	6.96
Move slowly	417	6.41
Swing	243	3.73
Boom down and raise the load	393	6.04
Boom up and lower the load	459	7.05
Stop	423	6.50
Emergency stop	489	7.51
Travel	237	3.64
Dog everything	285	4.38
Travel both tracks	246	3.78
Travel one track	288	4.43
Telescope out	333	5.12
Telescope in	303	4.66

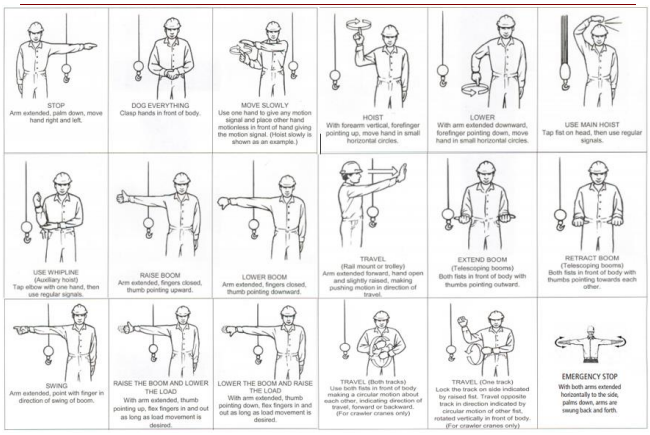

Figure 3. Crane signalman hand signals (Occupational Safety and Health Administration., 2009).

3.1.1. Data augmentation

Data augmentation is mandatory to better generalize the dataset, and it will help to increase the number of images in the dataset, which decreases the chances of model overfitting. There are several techniques that can be used to increase the size of a dataset and generalize the dataset, such as the cropping, rotating or flipping of the images. In the present work, the intensity transformation technique is used, which is an adjustment of the contrast and brightness of the images to generalize the image dataset for different scenarios like low and high light conditions. The data augmentation Increased the size of dataset by 25%.

3.2. Convolutional neural network-based deep learning model

3.2.1. *The architecture of the developed deep convolutional neural network*

Convolution neural networks are constructed by artificial neurons (i-e. mathematical functions that receive one or many inputs and sum them to produce an output) with weight, biases, and activation functions, which are responsible for transforming input images into a single output value. According to (LeCun and Bengio, 1995) Convolution neural networks use spatial decomposition of input images in multiple stages. Spatial decomposition is achieved through convolution and pooling layers. A DCNN is composed of four main features: convolution layer that is responsible to transform images through various filters to extract the features from the input, activation that adds non-linearity to the output

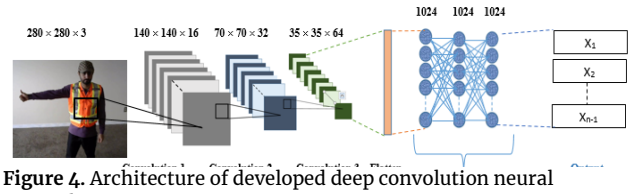

Figure 4. Architecture of developed deep convolution neural network

neurons, pooling or sub-sampling responsible for reducing dimensions of the feature map, and classification responsible for transferring output into a classification score.

The DCNN model developed for the proposed approach consists of convolution layers, pooling layers, dense layers, and an output layer. The input image in the DCNN has three color channels, red, green and blue (RGB), and it can be viewed as three 2-dimensional matrices arranged over each other having a pixel value in the range of 0 to 255 and is passed through convolution, pooling and dense layers to achieve an output vector. The architecture of the DCNN is shown in Figure 4.

The weights in the network are adjusted and optimized through the process of backpropagation. The process of backpropagation is achieved through epochs/ iterations such that the convolution neural network can correctly classify all images in the dataset. The architecture of the developed DCNN model consists of layers that are discussed in the following sections.

3.2.2. *Convolutional layers*

Convolution layers are considered the building blocks of convolution neural networks. The convolution operation is responsible for extracting features from the input image. The input image is convolved into a number of kernels. Kernels are the matrices used to store the weights of convolution operations.

Developed DCNN model architecture has three convolution layers. In the first convolution layer, the images are of size 280×280×3 (width, height, color channels) and each input image is convolved into 16 different kernels. Each kernel has a size of 3×3×3 (width, height, color channels). After the first convolution, each output has a size of 280×280 and 16 channels, so the resulting output will become 280×280×16. The resulting output is passed through the activation function and then subsampled to size 140×140×16 using max-pooling layers.

Second convolution layer takes the output of the max-pooling layer of size 140×140×16, which is further convolved with 32 different kernels, each kernel with a size of 3×3×16. This will result in 32 output channels of size 140×140. The resulting output is passed through the activation function and then subsampled to size 70×70×32 using max-pooling layers.

Third and final convolution layer takes the output of max-pooling after the second convolution layer of size 70×70×32 and is then convolved with 64 different kernels, each kernel with the size of 3×3×32. This will result in 64 output channels of size 70×70. After adding biases to 64 channels, the resulting output is passed through the activation function and then subsampled to size 35×35×64 using max-pooling layers.

The variation in the number of kernels is used in the architecture to obtain the highest accuracy in the model. Finally, a deep convolution neural network architecture with 16, 32, and 64 kernels in the first, second, and third convolution layers, respectively, produced the highest accuracy on the validation data set.

The developed network used a stride value of 2 that helps the kernel to move two matrix pixels at a time. This parameter affects the dimensions of output and reduces the chances of model overfitting. In the model, padding is used to assist the kernel to move uniformly over the matrix and its edges to obtain all the desired information in the image.

3.2.3. *Activation function*

Activation function is used in the convolution neural network to add non-linearity to the output neurons. Adding an activation function is essential, otherwise, the DCNN would compute linear combinations of linear functions and the model would not be able to learn complicated or non-linear functions (Nair and Hinton., 2010).

Activation function used in the developed DCNN is rectified linear unit (ReLU) and softmax. The ReLU activation function is an identity line where y=x for all positive lines and 0 for all negative values (Nair and Hinton, 2010). The mathematical equation for ReLU is given in equation 1.

$$f(x) = \max(0, x) \tag{1}$$

For the activation of the output layer, the softmax activation function is used. The softmax function is generally used for multiclass classification. It squashes the output of each unit to be between 0 and 1 and returns the probabilities of the input being in a particular class. Mathematically, it can be written as shown in equation 2.

$$P(yi\ \ I\ xi) = \frac{e^{f_{yi}}}{\sum_j e^{f_{yi}}} \tag{2}$$

Where yi = correct label of image xi, and f_{yi} = predicted score.

3.2.4. *Max-pooling layer*

Convolution layer along with the activation function is followed by the pooling or subsampling layer. The purpose of adding the pooling layer is to reduce the dimensions of the feature map and retain the important information (LeCun and Bengio, 1995). This layer also reduces computation and helps in reducing the overfitting of the model. In max-pooling, a kernel of size n × n is moved across the matrix and for each

position; the maximum value is taken.

In the developed DCNN, each convolution layer is followed by a max-pooling layer to reduce the dimension by a factor of 2. In the first convolution layer, max-pooling reduces the output channel from 280×280 to 140×140. In the second convolution layer, it further reduces the output channel from 140×140 to 70×70, and at the final convolution layer, the output channel is reduced from 70×70 to 35×35.

3.2.5. *Dense/fully connected layers*

Output of max-pooling layers is the input to the dense layer. In the dense layer, all input and output are connected to all the neurons in each layer, while neurons within a single layer share no connection. In a convolution neural network, dense layers are used to create the final non-linear combination of features and to predict the output layer. In the present study, three dense layers with 1,024 neurons in each layer are used. These layers are selected based on the classification performance of the validation set. A different number of dense layers is used and the accuracy of each is recorded. While recording the accuracy, it was noted that three dense layers with 1,024 neurons in each layer increase the average classification accuracy by 3%.

3.2.6. *Output layer*

Final layer of convolution neural network architecture is the output layer, which is responsible for transferring the output into a classification score. The softmax function is used for the activation of the output layer. The softmax function takes as input the predicted class labels and outputs a probability score. The equation for the softmax function is given in equation 2. The probability scores of output must have a sum of 1. The probability score indicates class prediction. The largest probability score for any hand signal is considered to belong to the correct class.

4. Model optimization

When the model achieves higher accuracy in the training dataset than the validation dataset, the model is considered to be overfitted. To prevent overfitting in the model, data augmentation is performed for the dataset (Chatfield et al. 2014), which includes brightness change, contrast change, image resizing, and image rescaling of images. Another method used to mitigate the overfitting in the model is dropout regularization. The dropout technique was proposed by (Srivastava et al., 2014). This technique activates the neurons with a certain probability and is implemented during the training stage. As the network is trained, neurons get randomly deactivated with respect to their weights and it will lead to better generalization of predictive capabilities (Srivastava et al., 2014). In the developed DCNN, a drop rate of 0.2 is chosen. A drop rate of 0.2 is selected based on multiple trails on the model, which gives higher accuracy during the validation and testing of the dataset.

5. Preparation of training, validation and test datasets

The dataset contains 8133 augmented and non-augmented images with 18 different crane signalman hand signals. The images were collected with individuals from the research team. The dataset was randomly split into three different datasets: a training set, a validation set, and a test set. The training set includes 70% of the images, while the validation and test sets each include 15%. The reason for choosing a larger sample size for the training set is to get more features from the training dataset, which leads to better accuracy in the validation and test datasets.

6. Model performance evaluation

The DCNN model is developed using Python with Keras and TensorFlow API, which is an open-source computer software library for dataflow. The training set is converged into 160 epochs (iterations) which were completed in 8 hours on windows operating system with Intel corei7 processor and GeForce RTX 3050ti graphics card. The number of epochs, 160, was chosen to achieve maximum accuracy in the validation dataset. As shown in Figure 5, the accuracies of the training and validation dataset increase when the number of epochs in the model is increased while the value of loss decreases with the increase of the number of epochs.

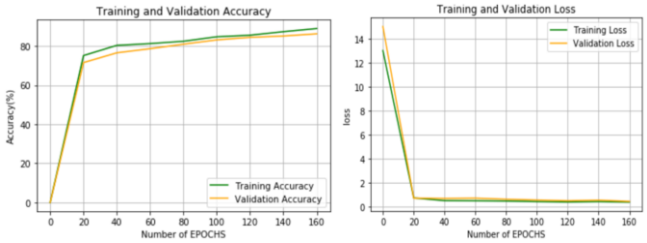

Figure 5. Training and validation accuracy and loss

The average training and validation accuracies achieved by the developed model were 89.1% and 84.6%, respectively and the loss calculated for training and validation was 0.38 and 0.44 respectively. The model calculates the accuracy using equation 3 and cross entropy loss using equation 4.

$$Accuracy = \frac{Correctly\ classified\ hand\ signal}{Total\ number\ of\ hand\ signal\ shown} \times 100 \quad (3)$$

$$cross\ entropy\ loss = \frac{-1}{N} \times \sum_{x=1}^{N} \sum_{y=1}^{M} Z_{xy} \times \log\left(p_{xy}\right) \quad (4)$$

Where N is the number of samples and M is the number of classes; Z_{xy} represents whether sample x belongs to class y or not and p_{xy} shows the probability of sample x belonging to class y. The loss has no upper limit and it exists in the range [0,∞]. The value of loss nearer to 0 indicates higher accuracy and vice versa.

Furthermore, a confusion matrix is used as a metric to evaluate the performance of the developed model on the test dataset. On the basis of the results obtained in the confusion matrix in the test dataset, the precision, recall, and F1 score are calculated using equations 5, 6, and 7.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Where TP is the number of true positives (the detected hand signal belongs to the class that is shown); FN is the number of false negatives which represent that the detected hand signal is not of the same class as actually shown, and FP is the number of false positives that represent that the detected hand signal is of different class and shown hand signal is of a different class. The model achieved an average precision for all crane signalman hand signals of 81.5%, and the average value recorded for the recall was 81.8%.

The other metric used to measure the performance of the model is the F1 score, which captures the properties of both precision and recall and combines them into a single unit. The reason for using the F1 score is that the model cannot be judged only on the basis of good results in either precision or recall. The F1 score for the developed model was recorded as 81.7%. Typically, the F1 score falls in a range from 0% to 100%, where 0% is poor performance and 100% is the best performance of the model.

6.1. Real-time classification

The model is further validated by deploying the developed model such that it is used to classify the crane signalman hand signals in real-time using a live stream. The developed DCNN model is validated in real-time using live stream by showing all the crane signalman hand signals a total of 802 times. The model was capable of correctly recognizing the hand signals 706 times with an average accuracy of 87.9%. The deep learning model was capable of correctly recognizing 42 hand signals out of 45 with an accuracy of 93.3%, 46 hand signals out of 48 with an accuracy of 95.8%, and 51 hand signals out of 52 with an accuracy of 98.1% for the labels hoist, stop, and emergency stop, respectively. The correctly and incorrectly recognized hand signals along with the accuracy can be seen in Table 2. The accuracy is measured using equation 3.

Standard Hand signals	Number of times crane signalman hand signal shown in camera	Number of times crane signalman hand signal is correctly classified	Number of times crane signalman hand signal is incorrectly classified	Accuracy (%)
Hoist	45	42	3	93.3
Lower	42	39	5	88.1
Use main hoist	49	39	10	79.6
Use whipline	46	40	6	87.0
Boom up	48	42	6	87.5
Boom down	45	39	6	86.7
Move slowly	48	44	4	91.4
Swing	47	38	9	80.9
Boom down and raise the load	40	35	5	87.5
Boom up and lower the load	39	34	5	87.2
Stop	48	46	2	95.8
Emergency stop	52	51	1	98.1
Travel	41	33	8	80.5
Dog everything	49	44	5	89.8
Travel both tracks	36	31	5	86.1
Travel one track	32	28	4	87.5
Telescope out	48	42	6	87.5
Telescope in	47	41	6	87.2

The proposed developed framework with deep learning model can be implemented in cranes using a camera and a screen/ head-up display. The camera is used to record the live stream of the signalman hand signals, which are transmitted to the screen placed inside the crane operator cabin. The developed DCNN model shows the results i-e. Classified labels of detected signalman hand signals to the screen. The operator inside the cabin can see the signalman hand signals and their classified labels on the screen. This developed framework assists the crane operator to take the decision about the movement of load more confidently and efficiently. In this way, the framework can be used as an improvement to the current state of practice for communication between crane operator and signalman and serve as another layer of communication and safety in crane industry.

Table 2. Accuracy of real-time crane signalman hand signal classification

7. Conclusions, limitations and future work

This developed framework presented in this paper used a DCNN model to classify the crane signalman hand signals in real-time, which will help to improve communication between crane operator and signalman. The DCNN model is trained using images of 18 hand signals that are used by the signalman to communicate instructions to the crane operator. The collected images are resized, rescaled, and the contrast and brightness of the images are adjusted to increase the number of images in the dataset, which leads to a better generalization of images in the dataset and decreases the chances of model overfitting. The 8133 images of the 18 crane signalman hand signals are passed through the deep convolution neural network to train the model to recognize the hand signals correctly. The architecture of the proposed DCNN model consists of 3 convolution layers, 3 max-pooling layers, 3 dense layers, and an output layer. The developed model achieved an accuracy of 89.1% and 84.6% in training and validation, respectively. The precision, recall and F1 score achieved by the model were 81.5%, 81.8%, and 81.7% respectively, for the test set. The model is further validated for real-time hand signals classification, where accuracy of 87.9% is recorded. The F1 score was greater than 80% which means the model is performing well and the average accuracy of 87.9% in real-time crane signalman hand signals classification makes the model acceptable. However, during real-time classification, some misjudgments are made by the DCNN model while classifying the signalman hand signals, but the crane operator is not relying solely on the classification label because the screen/ head-up display inside the operator cabin is showing the signalman hand signals in real-time along with the classification label. The crane operator can still take the correct action by looking at the signalman hand signal.

In terms of future research, there is room for improvement. For example, a larger dataset of crane signalman hand signals images can be used to train the model, which would lead to an improvement in the performance of the developed DCNN model. Another technique is the use of transfer learning with a pre-trained model in place of using the current model, which can reduce the computational time and increase accuracy in terms of correctly classifying the crane signalman hand signals. Data fusion techniques can also be used to classify dynamic hand signals that will lead to an improved the accuracy of the deep learning models. The development of such model is necessary in the crane industry as they help in the construction safety improvement and with well-achieved high accuracy; they assist towards automation of the process and development of automated cranes.

References

Arabi, S., Haghighat, A., and Sharma, A. (2019). "A deep learning based solution for construction equipment detection: from development to deployment." http://arxiv.org/abs/1904.09021.

Cha, Y. J., Choi, W., and Büyüköztürk, O. (2017). "Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks." *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361–378.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). "Return of the devil in the details: Delving deep into convolutional nets." *BMVC 2014 - Proceedings of the British Machine Vision Conference 2014*, 1–11.

Everett, B. J. G., and Slocum, A. H. (1993). "Device for Improving Crane." *Journal of Construction Engineering and Management*, 119(1), 23–39.

Fang, Y., and Cho, Y. K. (2016). "A framework of lift virtual prototyping (LVP) approach for crane safety planning." *ISARC 2016 - 33rd International Symposium on Automation and Robotics in Construction*, 291–297.

Gu, Y., Xu, S., Wang, Y., and Shi, L. (2019). "An advanced deep learning approach for safety helmet wearing detection." *Proceedings - 2019 IEEE International Congress on Cybermatics: 12th IEEE International Conference on Internet of Things, 15th IEEE International Conference on Green Computing and Communications, 12th IEEE International Conference on Cyber, Physical and So*, 669–674.

Hu, J., Geo, X., Wu, H., and Gao, S. (2019). "Detection of Workers Without the Helments in Videos Based on YOLO V3." *12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 1553–1560.

LeCun, Y., and Bengio, Y. (1995). "Convolutional Networks for Images, Speech, and Time-Series." *Handb. brain Theory Neural Netw.* 3361 (10).

Li, H., Chan, G., and Skitmore, M. (2013). "Integrating real time positioning systems to improve blind lifting and loading crane operations." *Construction Management and Economics*, 31(6), 596–605.

Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., & Al-Hussein, M. (2020). Conceptual Framework for Safety Improvement in Mobile Cranes. In *Construction Research Congress 2020: Computer Applications* (pp. 964-971). Reston, VA: American Society of Civil Engineers.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In: Proceedings of the International Conference on Machine Learning (ICML) 807–814

Occupational Safety and Health Administration (OSHA), Cranes and Derricks in Construction (1926.1408), https://open.alberta.ca/dataset/757fed78-8793-40bb-a920-6f000853172b/resource/9296e033-fd12-40dc-ac86-21e5873d4161/download/4403880-part-6-cranes-hoists-and-lifting-devices.pdf (2009) (accessed December 19, 2021)

Rosenfeld, Y. (1995). "Automation of existing cranes: from concept to prototype." *Automation in Construction*, 4(2), 125–138.

Shapira, A., Rosenfeld, Y., and Mizrahi, I. (2008). "Vision system for tower cranes." *Journal of Construction Engineering and Management*, 134(5), 320–332.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958

Wang, L., Xie, L., Yang, P., Deng, Q., Du, S., and Xu, L. (2020). "Hardhat-wearing detection based on a lightweight convolutional neural network with multi-scale features and a top-down module." *Sensors (Switzerland)*, 20(7), 3–7.

Wu, J., Cai, N., Chen, W., Wang, H., and Wang, G. (2019). "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset." *Automation in Construction*, Elsevier, 106.

Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., and Kurach, L. (2020). "A deep learning-based framework for an automated defect detection system for sewer pipes." *Automation in Construction*, Elsevier, 109(2019), 102967.

Zavichi, A., and Behzadan, A. H. (2011). "A Real Time Decision Support System for Enhanced Crane Operations in Construction and Manufacturing." *Computing in Civil Engineering*, 194–201.

Zekavat, P. R., Moon, S., and Bernold, L. E. (2014). "Holonic construction management: Unified framework for ICT-supported process control." *Journal of Management in Engineering*, 31(1), 1–15.