Classifying Excavator Operations with Fusion Network of Multi-modal Deep Learning Models

Jin-Young Kim and Sung-Bae Cho

Department of Computer Science, Yonsei University, Seoul, South Korea
(seago0828,sbcho)@yonsei.ac.kr

Abstract. Prognostics and health management (PHM) aims to offer comprehensive solutions for managing equipment health. Classifying the excavator operations plays an important role in measuring the lifetime, which is one of the tasks in PHM because the effect on the lifetime depends on the operations performed by the excavator. Several researchers have struggled with classifying the operations with either sensor or video data, but most of them have difficulties with the use of single modal data only, the surrounding environment, and the exclusive feature extraction for the data in different domains. In this paper, we propose a fusion network that classifies the excavator operations with multi-modal deep learning models. Trained are multiple classifiers with specific type of data, where feature extractors are reused to place at the front of the fusion network. The proposed fusion network combines a video-based model and a sensor-based model based on deep learning. To evaluate the performance of the proposed method, experiments are conducted with the data collected from real construction workplace. The proposed method yields the accuracy of 98.48% which is higher than conventional methods, and the multi-modal deep learning models can complement each other in terms of precision, recall, and F1-score.

Keywords: Excavator · Classification · Deep learning · Multi-modal data · Autoencoder · Feature extraction

1 Introduction

Remaining useful life (RUL) prediction is an attempt to predict the time period over which a device can perform normal operation at a certain level [1]. This is an important step that allows us to keep track of the expected lifetime of the equipment, or the time of replacement or repair, without interruption. Because we can expect the burden on the equipment depending on the mode of operation the equipment is performing, we can predict the RUL if we look at the operation history of the equipment. However, the process of manually identifying and classifying work histories (e.g., sensor record, survey record, etc.) is costly. Therefore, it is necessary to study the operation mode, and classify the operation mode accurately.

We can classify the operation mode of the equipment through the sensor log of the equipment or the video that shows the operation of the equipment. As direct viewing and classifying of videos can be highly accurate, the method of automatic classification...
of video has been studied extensively. However, as shown in Fig. 1, the performance varies greatly depending on the surrounding environment (weather, time, etc.) [2–4]. On the other hand, classifying the operation mode using sensor values is not significantly affected by the surrounding environment. However, as shown in Fig. 2, classification model using sensor values can hardly distinguish when the instrument performs similar operations. To show that it is difficult to classify the operation mode using the sensor data of the excavation used in this paper, the average and variance of the sensor values for each operation are illustrated in Fig. 3. Even though the statistics of the sensor data do not differ significantly from one operation to another, it is difficult to classify the operation of the equipment by using single-modal data. If we use the video and sensor values simultaneously, we can construct a model with higher performance and robustness.

![Fig. 1. Examples of video data. From left to right, images show the operations of leveling, excavation, rock excavation, and run.](image)

![Fig. 2. A sensor graph used in this paper. The sensor values when the equipment excavates the ground (left) and rock (right). X- and y-axes represent time and sensor values, respectively.](image)

In this paper, we propose a method to classify the operation mode by using both data at the same time to overcome the limitation of using one kind of data. Figure 4
Fig. 3. Statistics of sensor values. X- and y-axes represent the average and standard deviation for each operation, respectively.

shows some examples of operation mode of excavator. The method is based on deep learning models to effectively extract the characteristics of each data [5–8]. The classification model for each data is constructed and learned first, and the output of the last hidden layer of each model is set as extracted features, which are concatenated into the input of the final classification model to classify the operation mode. The main contributions of this paper are as follows.

• We construct a model that extracts the characteristics of multi-modal data (i.e., video and sensor data) which are different kinds of data, design a fusion network that integrates the features, and evaluate the performance with several experiments.
• We collect the data used in the real workplace and use it in experiments to show the practicality of the proposed model.

The paper is organized as follows. In Sect. 2, we present relevant works that solve the classification problem with one kind of data. We propose a method in Sect. 3 that classifies the operation mode with multi-modal data. Section 4 shows the experimental results to evaluate the proposed method. The conclusions and discussion of this paper are covered in Sect. 5.

Fig. 4. Examples of operation mode of excavator. Excavation is the digging operation and leveling is the operation to level the ground.
2 Related Work

In order to classify the operation mode of equipment by video data or sensor information, a method of processing each data is needed. Many researchers have carried out to perform tasks with video or sensor information, as summarized in Table 1.

In the sensor-based model, various feature extraction techniques are used. Dao et al. performed multi-sensor classification using a sparse representation framework [9]. Chavez-Garcia and Aycard performed tasks using localization and mapping, detecting and tracking the moving objects [10]. Cao et al. classified the excavation equipment by machine learning method after preprocessing sensor signals with mel-frequency cepstral coefficient (MFCC) technique [11, 12]. Choi and Cho constructed modular Bayesian networks and an optimal stimulus decision module for predicting emotion with various sensor information [13]. Kim and Cho studied a method to predict energy demand by processing power demand [14].

In the study on video-based model, many attempts have been made to effectively classify and use various features in video [2–4]. Donahue et al. performed frame-by-frame feature extraction based on the convolutional neural network and sequential information processing through recurrent neural network [15]. Zha et al. extracted features of each image frame and classified them with support vector machine (SVM) [16]. Ye et al. used a method of extracting features by applying not only video frames but also optical flow [17]. Wu et al. proposed a method to separate motion and background in video to classify it [18].

<table>
<thead>
<tr>
<th>Category</th>
<th>Author</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor-based model</td>
<td>Dao et al. [9]</td>
<td>Propose sparse representation framework</td>
</tr>
<tr>
<td></td>
<td>Chavez-Garcia and Aycard [10]</td>
<td>Use localization, mapping, detecting and tracking the moving objects</td>
</tr>
<tr>
<td></td>
<td>Cao et al. [11]</td>
<td>Present extreme learning model to classify the excavator</td>
</tr>
<tr>
<td></td>
<td>Choi and Cho [13]</td>
<td>Construct modular Bayesian networks to predict emotions in physical space with various sensor information</td>
</tr>
<tr>
<td>Video-based model</td>
<td>Donahue et al. [15]</td>
<td>Feature extraction by using convolutional neural network</td>
</tr>
<tr>
<td></td>
<td>Zha et al. [16]</td>
<td>Extract feature of each frame and classify video using SVM</td>
</tr>
<tr>
<td></td>
<td>Ye et al. [17]</td>
<td>Use video frame and extract feature through optical flow technique</td>
</tr>
<tr>
<td></td>
<td>Wu et al. [18]</td>
<td>Separate motion and background features to classify video</td>
</tr>
</tbody>
</table>
3 Proposed Method

The overall structure of the proposed method is shown in Fig. 5. It consists of a video-based model, a sensor-based model, and a fusion network that combines the two models. In order to extract the characteristics of video frames, we use an encoder, which is pre-trained with an autoencoder, in a video-based model. The features extracted for each frame enter the input of the time series model and are converted into the characteristics of the video data. The sensor-based model extracts the features by considering the various sensor information with the convolutional neural network, and then outputs the sensor data characteristics to the time series model. The fusion network concatenates the characteristics of each data obtained from the two models and uses them as inputs to classify the operation mode of the equipment.

![Fig. 5. The overall structure of the proposed model. It consists of a video-based model, a sensor-based model, and a fusion network. The video-based model extracts the frame feature with the encoder of the autoencoder and processes the sequence information to output the video feature. This is used as an input to the fusion-net with features extracted from the sensor-based model to make the model classify the operation mode.](image)

3.1 Video-Based Model

The video data can be viewed as image sequence. We first extract the features of each video frame and convert the image data into a sequence of spatial features. We use a model of an autoencoder to automatically extract image features. An autoencoder, a way to learn the representation of the data, consists of an encoder that compresses data and a decoder that reconstructs the data [14, 19, 20]. We use the frames of the task video to train the autoencoder. The objective function of autoencoder is as follows.

$$\mathcal{L}_{AE}^v = \mathcal{L}_r(g_v(f_v(x)), x),$$

(1)

where $\mathcal{L}_r : \mathcal{X} \times \mathcal{X}$ is a loss function measuring the reconstruction error from $g_v(f_v(x))$ to $x$, $g_v$ is a decoder, $f_v$ is an encoder, and $x$ is input data.

The encoder from learned autoencoder is reused in the video-based model to extract the features on a frame-by-frame basis. The sequence of extracted features is used as an input to the time series model and transformed into a feature of the video data. We use a long-short term memory (LSTM) network, one of the recurrent neural networks, as a
time series model [21]. In the video-based model, we train it to classify the operation of the equipment by adding a fully-connected layer after the time series model. The objective function of video-based model is shown in Eq. (2). However, when learning fusion network, we try to extract the features of video data by reusing the models except the fully-connected layer in the learned video-based model.

$$\mathcal{L}_v = \mathcal{L}_c(h_v(f_v(x)), y),$$  

(2)

where $\mathcal{L}_c : \mathcal{Y} \times \mathcal{Y}$ is a loss function measuring the difference between real label $y$ and computed output $h_v(f_v(x))$ and $h_v$ is newly added fully-connected layer.

### 3.2 Sensor-Based Model

We design a sensor-based model that extracts features by inputting 30 types of sensor values. All sensor values are preprocessed at the same interval by the sliding window technique and the features are extracted by the convolutional neural network. We also inspire the inception module to extract the characteristics of sensor values using various-sized filters [22]. The extracted features are processed by the LSTM model, and the operation mode of the equipment is classified by a fully-connected layer. As in the case of the video-based model, we try to extract the characteristics of the sensor data by reusing the models except the fully-connected layer in the learned sensor-based model when learning the fusion network. Unlike the video-based model, 1D convolutional neural network is used to extract the characteristics of sensor data. The objective function of sensor-based model is as follows.

$$\mathcal{L}_s = \mathcal{L}_c(h_s(f_s(x)), y),$$  

(3)

where $h_s$ and $f_s$ are a classifier and a feature extractor of sensor-based model, respectively.

### 3.3 Fusion Network

We propose a fusion network to extract the characteristics of multi-modal data and to classify the operation mode by integrating them. There are some methods to ensemble existing models, such as voting and averaging. However, as can be seen in Figs. 1 and 2, there are limits to conventional methods because the classes that can be classified according to the characteristics of one type of data are different. Therefore, we propose a fusion network that extracts the features of each data, reuses the feature extraction parts of models with different kinds of data, and performs classification by combining them. As depicted in Fig. 5, we concatenate features extracted from the video-based model and the sensor-based model, and use them as inputs to the new fully-connected layer. The newly added layers are trained to classify the operation mode by properly combining the extracted features with the weights of the new fully-connected layer. The objective function of fusion network is described as follows.
\[ \mathcal{L}_f = \mathcal{L}_c \left( h_f(f_v(x_v), f_s(x_s)), y \right), \]

where \( h_f : \mathcal{H} \times \mathcal{H} \) is a newly added fully-connected layer which gets extracted features of video and sensor data as input, \( \mathcal{H} \) is a space of extracted features, \( x_v \) is a video data, and \( x_s \) is a sensor data.

## 4 Experiments

### 4.1 Dataset and Experimental Settings

In order to verify the performance of the proposed model, we collected and used the video and sensor data of the excavator. Some examples of video data and sensor data used are shown in Figs. 1 and 2, respectively. A total of 30 sensor data are composed of 9 sensors associated with the engine, 12 sensors related to pressure, and 9 sensors associated with voltage and current. The frame size of the video data is \( 1920 \times 1080 \), and it is reduced to \( 320 \times 180 \) for computational convenience.

We performed a total of nine tasks and divided them into four classes. Information about each class is summarized in Table 2. We divided the collected data into 3:1 for training and test. The model distinguishes the operation type by looking at the operation video and sensor values for about 10 s.

**Table 2.** Details of data used in this paper. The “Time” column is the time of the collected operation mode, and the “Number” column is the number corresponding to the operation mode after the data preprocessing. The “Class” column represents the class that we want to classify.

<table>
<thead>
<tr>
<th>Operation mode</th>
<th>Time</th>
<th>Number</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leveling up-down</td>
<td>25’48”</td>
<td>1519</td>
<td>Leveling</td>
</tr>
<tr>
<td>Leveling front-back</td>
<td>33’12”</td>
<td>1963</td>
<td></td>
</tr>
<tr>
<td>Leveling left-right</td>
<td>25’26”</td>
<td>1497</td>
<td></td>
</tr>
<tr>
<td>Digging</td>
<td>27’20”</td>
<td>1604</td>
<td>Excavation</td>
</tr>
<tr>
<td>Deep excavation</td>
<td>54’40”</td>
<td>3218</td>
<td></td>
</tr>
<tr>
<td>Excavation</td>
<td>52’05”</td>
<td>3034</td>
<td></td>
</tr>
<tr>
<td>Slope excavation</td>
<td>28’33”</td>
<td>1666</td>
<td></td>
</tr>
<tr>
<td>Rock excavation</td>
<td>1:47’37”</td>
<td>6342</td>
<td>Rock excavation</td>
</tr>
<tr>
<td>Drive</td>
<td>58’12”</td>
<td>3463</td>
<td>Drive</td>
</tr>
</tbody>
</table>

### 4.2 Result Analysis

In this section, we illustrate the performance of the proposed method. We compare the 3D convolutional neural network, which has three-dimensional filters, Conv LSTM, which uses computation as a multiply operation in the typical LSTM, and the long-term recurrent convolutional networks (LRCN) model, which recurrently uses a convolutional neural network [23, 24]. As shown in Fig. 6, the proposed method has better performance than the conventional methods, as well as the models using one kind of data.
only. The performance of the fusion network is only 0.18% higher than that of the video-based model, but it has the advantage of complementing the advantages of both models.

We calculate the precision, recall, and F1-score for each class to see which operation the proposed model classifies well as shown in Table 3. The method solves the 4-class classification problem, but when calculating the precision, recall, and F1-score, the classification result is set as one versus rest for each class. In the leveling operation, the sensor-based model was stronger, and the video was better classified to distinguish the type of excavation. We can see the results of the fusion network taking advantage of both video and sensor-based model as a whole. Even in the excavation class, the performance exceeds that of the video and sensor-based models.

![Fig. 6. The experimental results of classifying the operation mode of excavator.](image)

**Table 3.** Results of classifying the operation with precision, recall and F1-score.

<table>
<thead>
<tr>
<th></th>
<th>Leveling</th>
<th>Excavation</th>
<th>Rock exc.</th>
<th>Run</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Video-based model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.9745</td>
<td>0.9813</td>
<td>0.9937</td>
<td>0.9838</td>
<td>0.9833</td>
</tr>
<tr>
<td>Recall</td>
<td>0.9758</td>
<td>0.9945</td>
<td>0.9994</td>
<td>0.9299</td>
<td>0.9749</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.9752</td>
<td>0.9878</td>
<td><strong>0.9965</strong></td>
<td><strong>0.9561</strong></td>
<td>0.9791</td>
</tr>
<tr>
<td><strong>Sensor-based model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.9723</td>
<td>0.9822</td>
<td>0.9226</td>
<td>0.9838</td>
<td>0.9652</td>
</tr>
<tr>
<td>Recall</td>
<td>0.9869</td>
<td>0.9540</td>
<td>0.9745</td>
<td>0.9299</td>
<td>0.9613</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.9796</td>
<td>0.9679</td>
<td>0.9478</td>
<td><strong>0.9561</strong></td>
<td>0.9633</td>
</tr>
<tr>
<td><strong>Fusion network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.9766</td>
<td>0.9862</td>
<td>0.9911</td>
<td>0.9828</td>
<td>0.9844</td>
</tr>
<tr>
<td>Recall</td>
<td>0.9830</td>
<td>0.9945</td>
<td>0.9994</td>
<td>0.9299</td>
<td>0.9767</td>
</tr>
<tr>
<td>F1-score</td>
<td><strong>0.9798</strong></td>
<td><strong>0.9903</strong></td>
<td>0.9952</td>
<td><strong>0.9561</strong></td>
<td><strong>0.9806</strong></td>
</tr>
</tbody>
</table>
5 Conclusion

In this paper, we addressed the need to classify the operation mode of excavator. Since there is a limit in classifying the operation mode with only sensor or video data, we constructed a method that performs tasks using both data. After learning the models that classify the excavator operation with each data, only the model that extracts the features was reused in the fusion network. The features extracted from the reused models enter the input of the new fully-connected layer and classify the operation of the excavator. The proposed method has the best performance compared to the conventional models with accuracy of 98.48%. Besides, through the analysis of F1-score, we confirm that the proposed method distinguishes each class complementarily from models using one data.

We will verify whether the reused feature extraction models extract the features of the video or sensor well by extracting the intermediate output. Although the disadvantages of sensor data that cannot distinguish between excavation and rock excavation are supplemented by the use of video data, we need to further collect various data to test whether the sensor data compensate for the disadvantages of the video data that are highly affected by the surrounding environment in the future. Moreover, we limited the number of classes to four, but we will classify all of the detailed classes in the future. Finally, we will apply the proposed model to an embedded board and install it in the real time classification system.

Acknowledgement. This work has been supported by a grant from Doosan Infracore, Inc.

References