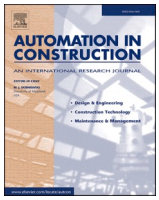
Automation in Construction 172 (2025) 106032

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon



# Automated recognition of construction worker activities using multimodal decision-level fusion

Yue Gong [a], JoonOh Seo [a,*], Kyung-Su Kang [b], Mengnan Shi [c]

[a] Department of Building and Real Estate, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[b] Department of Architecture, Sahmyook University, Seoul 01795, Republic of Korea
[c] College of Water Resource and Hydropower, Sichuan University, Chengdu, Sichuan 610065, China

ARTICLE INFO

*Keywords:*
Accelerometer
Activity recognition
Automation
Computer vision
Decision-level fusion
Dempster-Shafer theory
Multimodal fusion

ABSTRACT

This paper proposes an automated approach for construction worker activity recognition by integrating video and acceleration data, employing a decision-level fusion method that combines classification results from each data modality using the Dempster-Shafer Theory (DS). To address uneven sensor reliability, the Category-wise Weighted Dempster-Shafer (CWDS) approach is further proposed, estimating category-wise weights during training and embedding them into the fusion process. An experimental study with ten participants performing eight construction activities showed that models trained using DS and CWDS outperformed single-modal approaches, achieving accuracies of 91.8% and 95.6%, about 7% and 10% higher than those of vision-based and acceleration-based models, respectively. Category-wise improvements were also observed, indicating that the proposed multimodal fusion approaches result in a more robust and balanced model. These results highlight the effectiveness of integrating vision and accelerometer data through decision-level fusion to reduce uncertainty in multimodal data and leverage the strengths of single sensor-based approaches.

## 1. Introduction

The productivity of the construction industry is reported as one of the lowest across all industry sectors due to its heavy reliance on manual on-site operations [59]. To enhance labor productivity during on-site operations, practitioners have emphasized the need for field data collection not only to evaluate the productivity based on output per labor hour but also to measure the process of operations, aiming to understand the current status of site operations and identify causes of low productivity [28]. Traditionally, observation-based data collection has been adopted to systematically record both spatial (e.g., work zones) and temporal (e.g., task types, sequences, timings) information of on-site operations [25]. However, these methods have been criticized for being labor-intensive and time-consuming, leading to the urgent need for an automated means of field data collection [51].

Advancements in sensing technologies powered by data analytics techniques, such as machine learning, have facilitated the automated collection of field data to analyze the productivity of construction workers [34]. In particular, activity recognition using computer vision or wearable sensing technologies has been widely adopted in the construction domain as an effective means of measuring the temporal aspects of individual workers' tasks at a detailed level [37,75]. For example, construction tasks, such as concrete work, typically consist of a series of subtasks, including formwork, rebar placing, and concreting, which are further divided into specific activities such as transferring, positioning, or installing forms according to the work taxonomy [26]. Recognizing these activities from time series data, such as videos and acceleration data, enables the recording of sequences and durations for each activity, which can help identify productivity issues during manual operations. Vision-based activity recognition classifies predefined activities from image sequences, as different activities create distinguishable spatial and temporal features in images. Thus, machine learning algorithms can learn these features from training images to classify new images [9]. Similarly, wearable sensor-based approaches capture motion data, such as accelerations, from wearable sensors and classify activities by learning unique patterns of motion data associated with specific activities [50]. Both vision- and acceleration-based approaches have been validated in numerous studies within the construction field, demonstrating their reliability and applicability for evaluating worker-driven construction operations [7,36,46,74].

However, previous research efforts have pointed out the existing limitations of each approach, leading to the need for multimodal fusion

* Corresponding author.
*E-mail addresses:* yue-36.gong@polyu.edu.hk (Y. Gong), joonoh.seo@polyu.edu.hk (J. Seo).

https://doi.org/10.1016/j.autcon.2025.106032
Received 19 September 2024; Received in revised form 27 January 2025; Accepted 29 January 2025
Available online 7 February 2025

approaches [29]. Vision-based activity recognition has shown excellent performance in recognizing diverse construction activities even in harsh conditions such as dark or rainy construction sites, but its limited coverage due to camera positions and frequent occlusion issues has been widely criticized [11,65]. Additionally, distinguishing individual workers from images becomes problematic when multiple workers are present. In contrast, acceleration-based activity recognition is free of data loss and individual identification issues, as data is continuously collected from body-attached sensors. However, field tests in previous studies indicate that it tends to underperform compared with vision-based approaches, particularly due to confusion between activities with similar motions [26]. Recognizing this issue, previous studies in human activity recognition have introduced the fusion of images (primarily depth images) with inertial data, demonstrating more robust performance compared to single-modality approaches [69]. However, most algorithms only function when both image and acceleration data are available. This dependency may not be suitable for construction sites, where cameras often fail to capture workers who move out of the frame. Additionally, existing algorithms that attempt to address the limitations of single-modality approaches usually do not fully leverage the unique strengths of each modality in recognizing specific activities. For example, vision-based action recognition has demonstrated high accuracy for activities involving whole-body movements, such as laying bricks and tying rebars, but tends to struggle to recognize activities that involve fine hand movements like welding, sawing, and drilling [74]. Conversely, acceleration-based approaches can more accurately recognize hand-dominant activities when acceleration signals are captured from a wristband [26]. Therefore, fusing vision and acceleration data is expected to improve overall performance by compensating for the limitations of each modality's data.

This paper proposes an automatic approach for recognizing construction workers' activities by integrating video and acceleration data. The proposed method employs a multimodal decision-level fusion approach that combines local classification results from each modality using Dempster-Shafer Theory [78]. The originality of this research lies in the use of a decision-level fusion method and the adoption of category-wise weights for each modality's classification results in the fusion algorithm. The decision-level fusion allows classification results to be generated even in the absence of video data due to the limited range of cameras or occlusions, without any preprocessing or additional computational burden. Furthermore, the study leverages the complementary strengths of vision-based and acceleration-based recognition algorithms by introducing weighting strategies for fusion, informed by prior knowledge from model training. The decision-level fusion method is validated through experimental tests, comparing its activity recognition performance with those of single-modality approaches. The results provide insights into both the performance and the remaining challenges of the proposed method.

## 2. Literature review

### 2.1. Construction worker activity recognition using vision- and sensor-based approaches

Numerous studies have developed and validated vision- and sensor-based approaches for automated and efficient recognition of worker activities, as shown in Table 1. Vision-based methods show significant promise for recognizing construction worker activities due to the non-invasiveness of data collection and high recognition accuracy [43]. Multiple studies have validated the effectiveness of vision-based approaches for different trade workers, using various computer vision techniques and video datasets. For example, Liu et al. [45] proposed a silhouette-based human action recognition method using a single video camera to monitor worker activities on construction sites, achieving 90.7% accuracy in identifying predefined activity classes such as walking, lifting, and crawling in lab experiments. Also, Luo et al. [48] introduced a method for recognizing diverse construction activities in site images by leveraging convolutional neural networks (CNNs) to

**Table 1**
Vision- and acceleration-based method applications in construction worker activity recognition.

| Activity recognition method | Activity category | Classification accuracy | Research |
|---|---|---|---|
| Vision | Walking, tying rebar guiding crane, between activities | – | [12] |
| Vision | Fire caulking, hammering, idle, painting, walking | 85.3% | [24] |
| Vision | Breaking, cutting & measuring, holding, idling, picking up, putting down, walking | 76% | [37] |
| Vision | LayBrick, Transporting, CutPlate, Drilling, TieRebar, Nailing, Plastering, Shoveling, Bolting, Welding, Sawing | 61% | [74]; [73]. |
| Vision | Walking, lifting, crawling | 90.7% | [45] |
| Vision | Leveling land, Excavating for foundation, Placing concrete, Shipping materials, Finishing concrete, Installing foundation components, Transporting goods, Transporting people, Machining or transferring formwork, Building formwork of slabs and beams, Building formwork of walls and columns, Building formwork of stairs, Machining or transferring rebar, Fixing rebar of slabs and beams, Erecting rebar of walls and columns, Building scaffolding systems, Building scaffolding for slab formwork. | 62.4% precision and 87.3% | [48] |
| Vision | Collect plaster, Transfer plaster, Apply plaster, Prepare material, Place material, Consolidate placement | 78.5% | [53] |
| Vision | Driving truck, Transporting cement, Checking the power socket, Cleaning up the plank, Lashing rebar, Paving concrete, Installing scaffolding, Smearing plaster | 93.9% (non-occlusion) and 86.6% (under occlusion) | [44] |
| Acceleration | Work, material, travel, and idle | – | [17] |
| Acceleration | Effective work, essential contributory work, ineffective work | 90.1% (ironwork) and 77.7% (carpentry) | [36] |
| Acceleration | Loading, pushing, unloading, returning, idling | 87% to 97% (user-dependent) and 62% to 96% (user-independent) | [2] |
| Acceleration | Spreading mortar, laying blocks, adjusting blocks, removing mortar | 88.1% | [54] |
| Acceleration | Standing, bending-up, bending, bending-down, squatting-up, squatting, squatting-down, walking, twisting, working overhead, kneeling-up, kneeling, kneeling-down, and using stairs | 94.7% | [39] |

detect 22 classes of construction-related objects. This approach emphasizes the use of semantic relevance, representing the likelihood of cooperation or coexistence between objects, and spatial relevance, indicating pixel proximity, to recognize 17 types of construction activities. Similarly, Roberts et al. [53] developed a deep learning- and vision-based framework for activity analysis, specifically targeting construction resources through the analysis of 2D worker poses in RGB video footage. This paper demonstrated a high level of accuracy with 82.6% mean average precision for pose estimation and 78.5% cross-validation accuracy for activity analysis. Despite numerous previous studies showing the potential of vision-based approaches for the activity recognition of construction workers, challenges in capturing site images or videos with clear views have been frequently reported. These challenges are particularly prevalent in dynamic construction environments, where occlusions and poor lighting conditions often occur. For instance, Li and Li [44] evaluated the performance of a GAN-based model integrated with an attention mechanism for activity recognition from images with and without occlusions and found that occluded images could lead to almost a 10% drop in accuracy. Also, Hussain et al. [33] found that poor lighting conditions could result in a significant decrease in the performance of vision-based activity recognition systems, as traditional cameras are unable to function effectively in low-light or dark environments.

In the application of sensor-based activity recognition in the construction domain, accelerometers are among the most widely used sensors for automatically monitoring worker activity [26]. Compared with vision-based activity recognition approaches, acceleration-based methods typically utilize accelerometers or inertial measurement units (IMUs) embedded in wearable devices to capture acceleration signals from body movements. Thus, they are not affected by environmental conditions such as occlusions and varying lighting conditions [2]. Consequently, acceleration-based activity recognition has garnered significant attention for its potential to facilitate continuous and automated activity analysis in dynamic environments such as construction sites. A typical pipeline for acceleration-based methods involves recording acceleration data from construction activities, segmenting and labeling the signals through sliding window techniques, and training machine learning-based classification algorithms [72]. In the construction domain, the use of acceleration signals from wearable sensors or mobile phones with IMUs has been extensively investigated for various types of construction activities. For example, Joshua and Varghese [35] demonstrated that acceleration-based activity recognition could achieve about 80% classification accuracy for various masonry tasks, proving substantial promise for this approach for activity analysis. Akhavian and Behzadan [2] validated the effectiveness of using smartphone-collected acceleration signals to distinguish idling and sawing activities. Gong et al. [26] developed a hierarchical work taxonomy for acceleration-based activity recognition in construction, categorizing activities by body movements and work contexts to enhance classification accuracy. When tested with data from 18 workers, the approach achieved over 90% accuracy for Level 1 activities, 80–90% for Level 2, and around 75% for Level 3 by employing machine learning algorithms. Despite promising results, significant challenges remain in achieving better accuracy in worker activity recognition. For example, similar postures and movements involved in different activities (e.g., tying, screwing) tend to lead to confusion between them as they would create relatively similar acceleration patterns [1]. Another limitation of using acceleration data alone is the lack of sufficient semantic information to recognize complex activities accurately [52]. For instance, activities like walking without carrying any objects and transferring materials are distinct in their work context, yet they may generate similar acceleration signals. This similarity makes it challenging to differentiate between these activities based solely on acceleration data.

In summary, both vision- and acceleration-based activity recognition have shown their potential for capturing temporal (e.g., task types, sequences, timings) information of on-site operations that can be used for productivity analysis, especially for worker-oriented manual operations. However, there are several remaining challenges for practical applications of these approaches. Even though recent studies have tried to address these challenges by proposing new methodologies or algorithms, some of the challenges (e.g., occlusions, limited coverage range of vision-based approaches, confusions between activities, or lack of contextual information of acceleration-based approaches) are inherent, which are not easy to address by a single modality approach. This problem has led to the need for a sensor-fusion approach for activity recognition that combines data from different sources not only to achieve better accuracy but also to reduce uncertainty. In particular, the fusion of vision and acceleration data has demonstrated advantages in improving classification accuracy across various domains, including gesture recognition, activity recognition [16], and infrastructure health monitoring [60,71]. Within the construction domain, Kim et al. [38] proposed a hybrid kinematic (i.e., acceleration signals) and visual sensing framework that integrates features from both sources to conduct activity recognition of construction equipment. However, the feasibility of applying this approach to worker activity analysis remains to be validated.

### 2.2. Multimodal fusion techniques and applications

Multimodal fusion, a technique that merges information from multiple sources, enhances tasks' accuracy, stability, and efficiency compared to single-modality data [23,49]. The advent of advanced sensor technologies has enabled efficient, precise, and automated data acquisition, resulting in the development of various fusion types [55,66]. Each form of fusion possesses strengths suited to specific tasks, and thus, selecting the appropriate fusion strategy is crucial. Researchers have devised multiple fusion architectures tailored to the unique attributes and practicality of these methods, thereby aiding in the effective identification of appropriate fusion techniques [4]. Luo and Kay [47] proposed a framework of fusion approaches that can be classified into three categories: data-level, feature-level, and decision-level fusions, according to the steps of data processing, as shown in Fig. 1. Data-level fusion necessitates that raw data from multiple sensor sources be combined, after which the integrated data serve as input for the single model (Fig. 1 (a)). For effective data fusion at this level, the raw data must be both commensurate and appropriately associated prior to the fusion process [14]. Meanwhile, as reported by [42], the computational cost associated with data-level fusion exceeds that of feature- and decision-level fusion methods because of its extensive data preprocessing. Feature-level fusion focuses on extracting and amalgamating feature vectors derived from each sensor's observations as a part of data processing in the model. The vectors are then coalesced into a singular and comprehensive feature vector that is processed by the model, such as

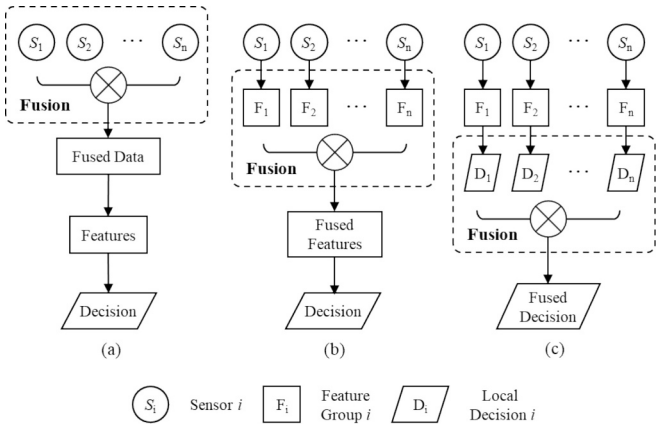



**Fig. 1.** Hierarchical data processing framework for fusion: (a) Data-level fusion, (b) Feature-level fusion, (c) Decision-level fusion.

**Table 2**
Comparison of data-, feature-, and decision-level fusion.

| Fusion type | Advantages | Disadvantages |
|---|---|---|
| Data-level fusion | Comprehensive data preservation | High computational power and data compatibility |
| Feature-level fusion | Dimensionality reduction and incompatible data tolerance | Critical feature choice and potential data loss |
| Decision-level fusion | Computational efficiency and high system flexibility | Lack of data detail |

neural networks, culminating in a unified output based on the amalgamated feature vectors from all sensors (Fig. 1 (b)) [31]. By employing feature engineering strategies such as dimensionality reduction, the complexity and computational cost of the feature-level approach can potentially be reduced. Nonetheless, the feature-level approach has limitations, as it might involve data loss during the feature selection or filtering processes. This limitation could lead to a potential issue related to data integrity [6,68]. Decision-level fusion architecture requires each sensor to independently generate a preliminary result based on its specific data set and an independent model (Fig. 1 (c)). After that, the final decision is made by integrating these preliminary results using methods like classical inference, Bayesian inference, or Dempster-Shafer's theory [31]. The comparative merits and limitations of data-, feature-, and decision-level fusion are summarized in Table 2. Among the three types of fusion approaches, the decision-level fusion would be the most appropriate for recognizing construction workers' activities due to its robustness against data incompleteness and minimal requirement for complex data preprocessing. Also, considering that activity recognition algorithms using vision or acceleration data are rapidly advancing, the decision-level fusion approach would allow quick updates of new classification models without disrupting the existing sensor network. These advantages can enhance flexibility and practical applicability [13,16,30,67].

The performance of decision-level fusion heavily relies on the methods used to make final decisions based on the outputs from different sensor sources. In general, the majority voting method is widely used to consolidate outputs from various predictions, estimates, or classifications and to decide the final output based on the most frequently occurring result [58]. Majority voting is most suitable for binary or discrete decision problems (e.g., yes/no, true/false). It is not well-suited for continuous or nuanced decision-making where there may be a spectrum of possible outcomes or where the decision needs to reflect degrees of confidence [20]. To address this issue, multimodal fusion with Bayesian inference integrates multiple pieces of evidence using Bayesian probability theory. This approach represents the uncertainty of an event through conditional probabilities ranging from zero to one. The core principle of Bayesian fusion is the use of posterior probabilities to represent beliefs about the fusion results. In multisensory data fusion, decisions made by each sensor serve as prior probabilities for specific conditions. By integrating these probabilities, Bayesian fusion combines data from multiple sources to create an overall likelihood that supports a particular hypothesis. The final decision is made by identifying the outcome with the highest combined likelihood [61]. Zappi et al. [76] investigated activity recognition using on-body sensors during automotive manufacturing processes was investigated. Two methods were proposed to address challenges such as sensor degradation, interconnection failures, and variations in sensor placement and orientation: a naive Bayesian fusion technique and a majority voting scheme. The results showed that the naive Bayesian fusion method significantly enhanced classification accuracy, increasing it from 50% with a single sensor to 98% with 57 sensors. Furthermore, the comparison between the two methods demonstrated that the Bayesian approach outperformed the majority voting scheme in this paper.

Recently, many studies on decision-level fusion have been applying the Dempster-Shafer Theory (DST) as a final decision model. It is a mathematical theory in statistics that was first proposed by Dempster [19] and further developed by Shafer [57]. The method, also known as evidence theory or belief theory, introduces the belief function to quantify the degree of belief for a particular hypothesis based on the available evidence, allowing for uncertainty representation without necessitating the summation of probabilities to one across the sample space [27]. Unlike Bayesian inference, which requires predefined prior probabilities, DST does not rely on such information for decision fusion, thereby thereby eliminating biases from hand-crafted priors [18]. Its flexibility makes it a better choice for fusing results from various sensors in multimodal systems [13]. For instance, Chen et al. [15] applied a DST-based decision-level fusion method for human activity recognition using both a depth camera and an inertial sensor. Tests conducted on a public human action dataset showed that the decision-level fusion method outperformed each sensor used individually (e.g., Kinect, accelerometer).

When applying the Dempster-Shafer (DS) theory for decision-level fusion, a common assumption is that all sensors have equal credibility [22]. However, due to the relative strengths and weaknesses of vision and acceleration data in recognizing specific activities, one type of sensor may outperform the other depending on the activity. For example, acceleration signals from wristbands are likely to more accurately capture hand-dominant activities, such as tying rebar, while images may be less effective in detecting small hand movements, such as those involved in assembling small parts. To address this issue, previous studies have adopted a weighted method that applies a discounting variable to each fusion input, thereby adjusting for differing trust levels among sensor sources [57]. In general, the weighting factors would be determined based on expert knowledge [10,63] or historical data Wu et al. [70]. However, it would be necessary to identify the most suitable way to determine the weighting factors for vision- or acceleration-based activity recognition results, considering the nature of construction tasks and strengths of each activity recognition method.

## 3. Methodology

This paper proposes a multimodal fusion approach for automated activity recognition by combining RGB images from site videos with acceleration signals from a wristband. The objective is to measure key temporal information, such as the types, sequences, and durations of manual handling tasks in construction. The proposed approach employs a decision-level fusion method based on the Dempster-Shafer theory to determine activity classes from two independent deep learning models trained with vision and acceleration data, respectively. Category-wise weights, calculated during the training process to account for the unequal trust levels of each sensor, are applied in the final decision model. These features, combining decision-level fusion with category-specific weights, ensure predictions despite data loss and improve classification performance by maximizing the strengths of each modality. Fig. 2 shows an overview of the proposed methodological framework, with its details described in the following sections. This framework consists of four key components: 1) data pre-processing, 2) deep learning-based activity recognition, 3) category-wise weight extraction, and 4) decision-level fusion. The framework begins with the collection of construction activity data, including RGB video images from cameras positioned at sites and acceleration signals from wrist-worn activity trackers. The data are then pre-processed for segmentation and synchronization. Deep learning algorithms use these two data modalities to independently generate preliminary activity predictions. These initial predictions, which are the local decisions, are adjusted using the estimated category-specific weights and are subsequently integrated using

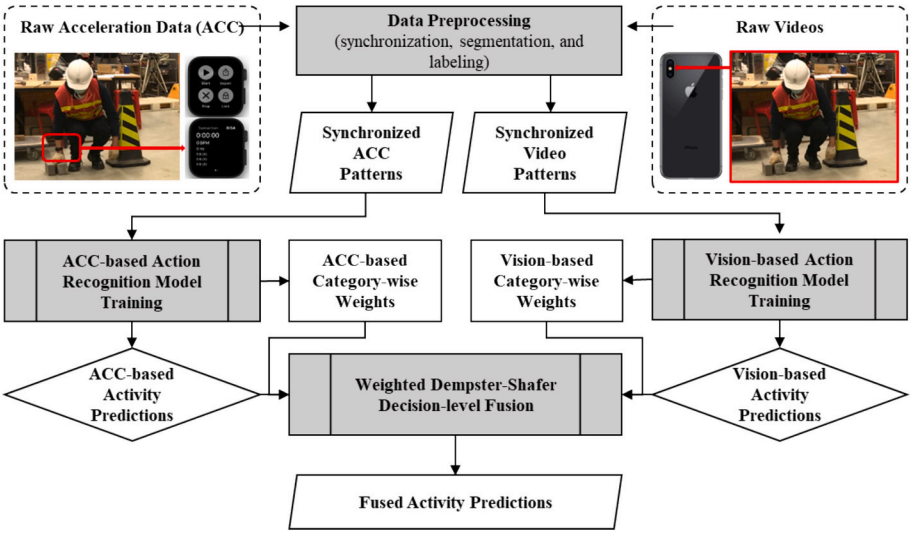



**Fig. 2.** Weighted decision-level fusion framework for worker activity recognition.

the Dempster-Shafer method. The effectiveness of this framework has been validated through experiments.

### 3.1. Data preprocessing

Data preprocessing involves three main steps: synchronization, segmentation, and labeling. First, video and acceleration data are synchronized to ensure temporal alignment and mitigate discrepancies [41]. Proper synchronization is essential to maintain temporal coherence, ensuring that data points from different sources represent the same moment, which is vital for accurate labeling and analysis [3]. Next, segmentation is applied, which is critical for Human Activity Recognition (HAR) tasks that use time-series data such as acceleration measurements [8,77]. This paper employed a sliding window technique to divide continuous acceleration data into smaller, fixed-size sequences [21]. Based on previous research, window sizes ranging from 0.5 to 4.0 s, in 0.5-s increments, were tested to identify the optimal size for maximizing classification performance, with a 50% overlap applied to minimize transition noise [54,62]. Although the vision-based activity recognition algorithm does not require segmentation, the video data were slid into sequences matching the acceleration segments to ensure consistency. Preliminary testing for different window sizes indicates that the optimal window size for the acceleration approach is 1.0 s, while the optimal window size for the vision approach is 0.5 s according to the average testing accuracy. The results indicate that the acceleration model is more sensitive to window size. Therefore, the video and acceleration data are both segmented using a 1.0-s window.

For labeling, this paper adopted a hierarchical activity taxonomy from a previous study [26]. The taxonomy consists of three levels: Level 1 distinguishes between “Idling” and “Work”; Level 2 further divides “Work” into “Traveling” and “Material Installation” (e.g., preparing, connecting, placing materials); and Level 3 provides more detailed subcategories such as “Material preparation” (e.g., cutting, bending), “Material connecting” (e.g., tying, screwing), and “Material placing” (e.g., lifting, adjusting). This hierarchical taxonomy facilitates comprehensive productivity analysis by identifying different activity types and specific areas of inefficiency. The entire preprocessing procedure is illustrated in Fig. 3, where the preprocessed data are prepared for input into single-modal deep learning models to recognize worker activity.

### 3.2. Single-modal deep learning models for construction worker activity recognition

The segmented datasets, labeled with predefined activities, were used to train two independent single-modal deep learning models, which form the basis for a multimodal fusion approach to activity recognition. For the vision-based model, the core architecture is ResNet (Residual Networks), a deep neural network design that effectively addresses degradation problems through residual mapping and shortcut connections [32]. Among the various ResNet variants, this paper utilizes ResNet-50, a 50-layer deep neural network architecture designed to efficiently learn and infer complex patterns in video data, which has been validated in prior research for activity recognition tasks [64]. In contrast, the acceleration-based model is trained using a Bidirectional Long Short-Term Memory (BiLSTM) network, an architecture well-validated for classifying diverse construction tasks using acceleration

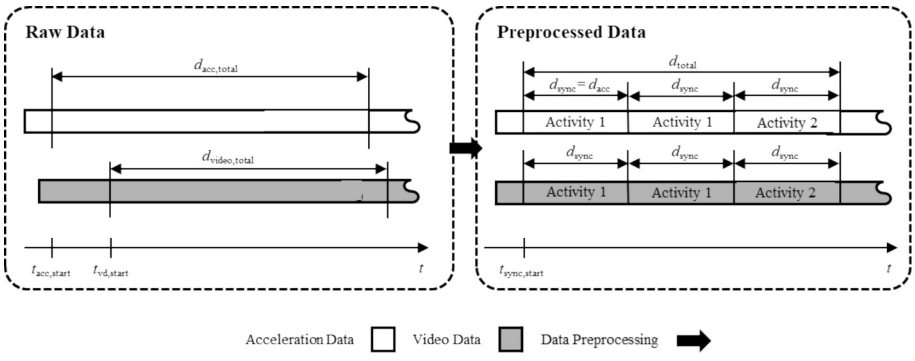



**Fig. 3.** Data preprocessing illustration.

signals [26].

The trained single-modal models produce a predicted activity label during inference, typically the one with the highest probability, which serves as a key metric for evaluating model performance. However, when combining predictions from multiple single-modal models, conflicts may arise due to differing predictions. To resolve these conflicts, the models fuse their confidence scores rather than their predicted labels. These confidence scores, generated by the softmax function in the final layer of neural networks, represent a probability distribution across all classes and quantify the network's confidence in its predictions [5,40]. Fig. 4 illustrates this process. The confidence scores are then used as input to the fusion module, enabling decision-level fusion by integrating local predictions from multiple models.

### 3.3. Decision-level fusion model for construction worker activity recognition

The decision-level fusion model makes a final determination based on the classification results from the vision- and acceleration-based activity recognition models developed in the previous step. The fusion model utilizes the Dempster–Shafer Theory (DST) to combine evidence. However, to address the issue of varying credibility between the sensor systems, the current study applies category-wise weights to the prediction scores to balance their respective influences. Fig. 5 illustrates the overall framework for the decision-level fusion.

Employing Dempster-Shafer Theory (DST) to fuse activity recognition results in decision-level fusion, which involves two steps. The first step converts the possibility of predicted labels to the degree of belief in the label. The second step combines the beliefs for the same class label obtained from the two activity recognition models using Dempster's rule of combination [13,56]. The belief degree is measured by the mass function, which has the following properties:

$$m(\varnothing) = 0 \tag{1}$$

$$m : 2^{\Theta} \rightarrow [0, 1] \tag{2}$$

$$\sum_{\{H \in \Theta\}} m(H) = 1 \tag{3}$$

where space $\Theta$ denotes all possible predicted classes, $H$ represents the predicted class by the model, and $m$ is the mass function. Additionally, the combination rule for DST is expressed by the following equations:

$$m_{\mathrm{v,a}}(H) = m_v \oplus m_a = \frac{1}{1-K} \sum_{\{B \cap C = H\}} m_{\mathrm{v}}(B) m_{\mathrm{a}}(C) \tag{4}$$

$$K = \sum_{\{B \cap C = \varnothing\}} m_{\mathrm{v}}(B) m_{\mathrm{a}}(C) \tag{5}$$

$$m_{\mathrm{v,a}}(\varnothing) = 0 \text{ and } H \neq \varnothing \tag{6}$$

Hypotheses $B$ and $C$ refer to the classification results of the vision-based model and the acceleration-based model, respectively. In this paper, the outcome of classification is a set of probability scores for potential activity categories rather than a single definitive outcome. For example, hypotheses $B$ = ["Drill": 0.26, "Hammer": 0.01, "Idle": 0.00, "LiftBrick": 0.05, "LiftRebar": 0.10, "MeasureRebar": 0.47, "TieRebar": 0.09, "Travel": 0.02], and corresponding hypotheses C could be represented as ["Drill": 0.61, "Hammer": 0.00, "Idle": 0.00, "LiftBrick": 0.00, "LiftRebar": 0.03, "MeasureRebar": 0.33, "TieRebar": 0.02, "Travel": 0.00]. The mass functions $m_{\mathrm{v}}(B)$ and $m_{\mathrm{a}}(C)$ are constructed based on the hypotheses $B$ and $C$, respectively. The combined belief degree $m_{\mathrm{v,\,a}}(H)$ represents the confidence that both models recognize the activity as category $H$. Additionally, the conflict measure $K$ quantifies the disagreement between the mass functions from the two different sources. In situations where one modality is missing, the method addresses this by treating the mass function of the missing modality as vacuous, which assigns all belief to the entire frame of discernment $\Theta$, in the Dempster-Shafer theory. In particular, a vacuous mass function meets the following principles:

$$m(B) = 0 \text{ for all } B \subset \Theta \text{ and } B \neq \varnothing \tag{7}$$

By introducing the above assumptions, the $B \cap C$ in Eq. (5) equals $C$ (the available modality's hypothesis) rather than the empty set, allowing the combination to proceed correctly. To achieve this, the output probabilities of the missing modality are set to zero for all categories, effectively constructing the vacuous mass function in the combination process.

While the Dempster-Shafer approach serves as a baseline, an additional weighting strategy is applied to assign category-specific weights to each model's predictions, ensuring balanced trust levels for different sensor sources. This modified method, called the Weighted Dempster-Shafer Theory (WDST), integrates the decision using the following updated combination rule:

$$m_{\mathrm{v,a}}(H) = m_v \oplus m_a = \frac{1}{1-K} w_v m_v(B) \times w_a m_v(C) \tag{8}$$

In the equation, $w_{\mathrm{v}}$ and $w_{\mathrm{a}}$ are the category-specific weight vectors for the vision- and acceleration-based activity recognition models, respectively. Selecting these weights is essential for ensuring the credibility and accuracy of the sensor systems. Traditionally, weights are determined using historical data or expert knowledge. However, this paper uses category-wise accuracy metrics obtained from model validation as weights for each activity prediction. This approach is a specific implementation of WDST and is denoted as Category-wise Weighted Dempster-Shafer (CWDS) in this paper. The proposed weight estimation method is an adaptive approach that reflects the model's category-wise potential. Whenever the dataset or algorithm changes, the estimated weights are updated accordingly, indicating a more reliable confidence level for the model's predictions in each category. Additionally, metrics such as precision, recall, and F1 score, which are used to evaluate the performance of classifiers, can be applied as weights. These metrics indicate each model's strength in predicting specific activities. By using these metrics as category-specific weights for each model's predictions, the fusion model can rely more on the model that demonstrates higher reliability in detecting a given activity class.

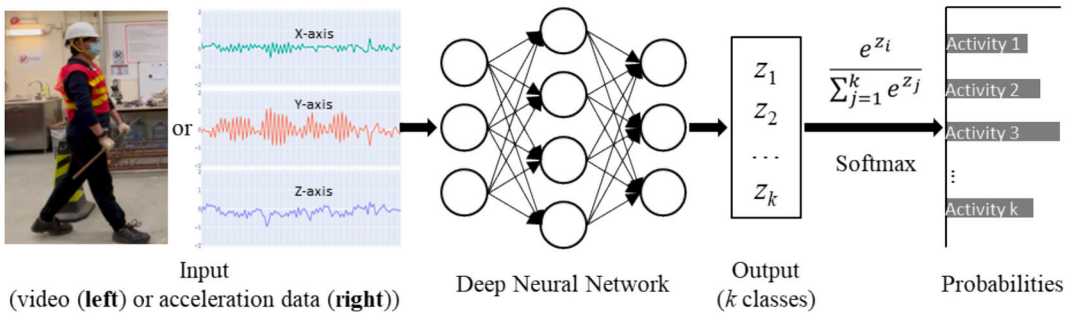



**Fig. 4.** Class probability distribution for multiclass activity recognition using softmax activation in deep learning.

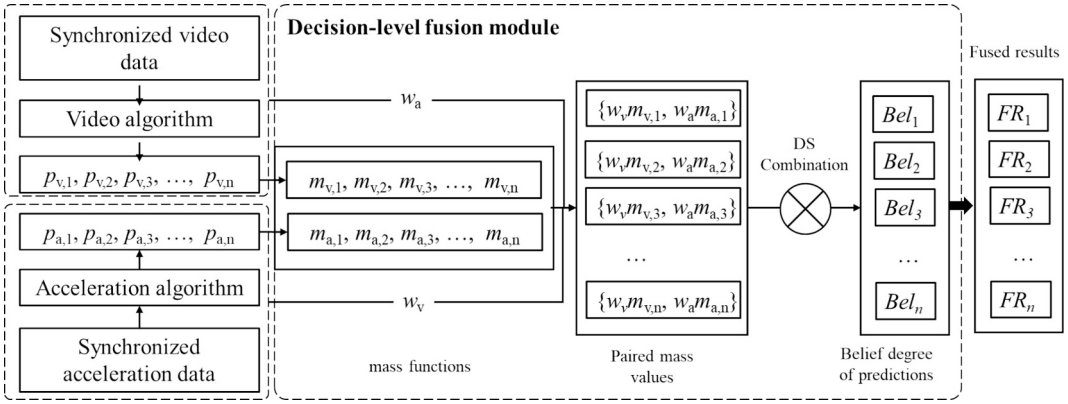



**Fig. 5.** Weighted Dempster-Shafer decision-level fusion framework.

In the practice of activity recognition model inference, many scores of categories towards extremely low values. For example, in one output possibility distribution, one category has 0.9 possibilities while the remaining categories have extremely small values, such as $10^{-6}$. These minimal values indicate a small likelihood of classification for those categories and have little impact on results when considering the conflict. However, when introducing the weight during the DS combination process, these extreme values introduce considerable errors to the combination results. The current study, therefore, proposed a strategy to filter out extreme predictions. The filtering approach involves sorting the prediction scores in descending order and filtering out weights below a specified threshold, such as 0.001. The remaining probabilities and their corresponding prediction categories are then re-normalized and constructed to the mass function. The subsequent fusion process adheres to the Dempster-Shafer rule of combination.

### 3.4. Experimental conditions and data processing

Experimental tests were conducted to validate the proposed method by simulating construction tasks and evaluating activity recognition performance using video and acceleration data. Acceleration data were collected from an Apple Watch (Series 4, 40 mm, 30.1 g) worn on the dominant arm. Simultaneously, videos were recorded from three different angles (i.e., front, side, and diagonal) using embedded smartphone cameras (iPhone X, iPhone 7, and iPhone 12). The sampling frequencies were 100 Hz for acceleration data and 30 fps for video data. Ten postgraduate students from the Hong Kong Polytechnic University, majoring in Building Engineering and Management, participated in the experiment. Each participant wore full safety gear, including hard hats and safety vests, to simulate real-world conditions. All participants gave their informed consent following the procedures approved by the Human Subject Ethics Subcommittee of the Hong Kong Polytechnic University (Reference Number: HSEARS20161102003). The experiments were conducted in laboratory settings designed to simulate construction environments without intentional occlusions.

The participants simulated eight typical construction activities in the laboratory (Fig. 6): “Traveling” (TL), “Lifting Brick” (LB), “Lifting Rebar” (LR), “Measuring Rebar” (MR), “Tying Rebar” (TR), “Hammering” (HR), “Drilling” (DR), and “Idling” (ID). These activities were selected to cover a range of body and hand movements. For example, “Traveling” and “Idling” involve whole-body horizontal movements, while the lifting tasks engage both whole-body and arm movements. The use of different materials, such as brick and rebar, introduces slight variations in arm motions due to the distinct handling requirements of each material. The remaining activities, such as “Measuring Rebar,” “Tying Rebar,” “Hammering,” and “Drilling,” represent material assembly tasks that require minimal whole-body movement and more precise hand and arm actions. To enhance data

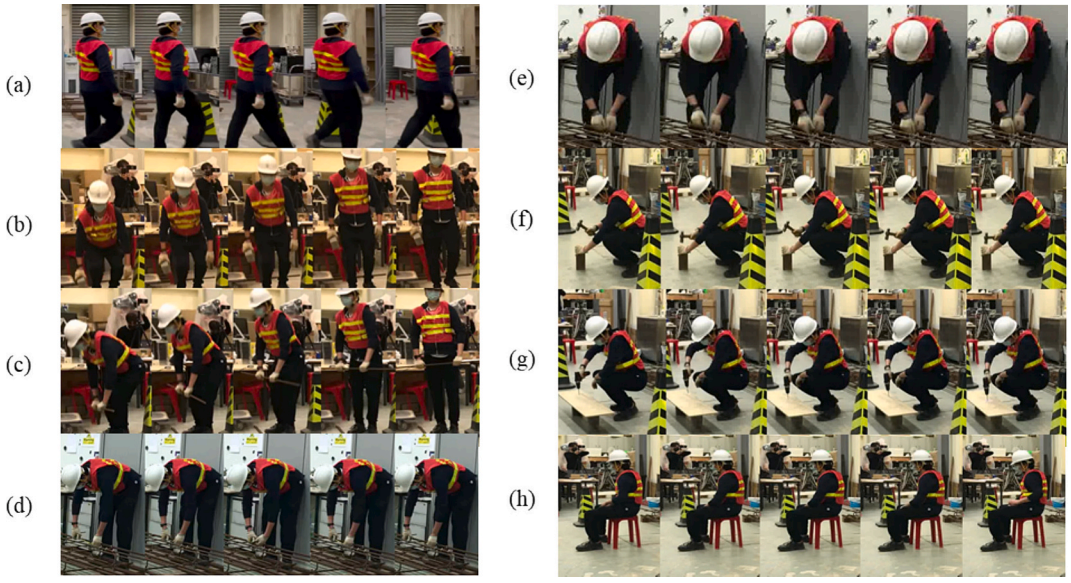



**Fig. 6.** Images of instructed activities: (a) Traveling, (b) Lifting Brick, (c) Lifting Rebar, (d) Measuring Rebar, (e) Tying Rebar, (f) Hammering, (g) Drilling, and (h) Idling.

reliability, minimize variability, and ensure comprehensive data collection, the researchers asked participants to repeat each activity five times, resulting in a dataset of 240 well-labeled trials. This experimental setup provides a robust foundation for developing and validating deep learning models for activity recognition. The controlled experimental conditions, characterized by limited inter-subject variability, aimed to validate the methodological framework for vision and sensor data fusion at the decision level.

Before starting the activities, participants were asked to perform a warm-up action by continuously waving their dominant hand while wearing the Apple Watch. This generated distinct acceleration signals that were accurately recorded and easily identifiable in both the acceleration data and the video footage. The moment the hand stopped waving served as a synchronization reference point, enabling precise alignment of timestamps between the videos from multiple camera angles and the acceleration data. After the warm-up concluded, the actual activity began and continued until the participant returned to a designated spot, as observed in the video recordings. The endpoint of the acceleration data was determined by matching the duration of the corresponding video event from the aligned start time. Following this synchronization, both video and acceleration data were segmented with a window size of 1.0 s, which demonstrated the best performance in classifying activities using each activity recognition algorithm. The segmented data were then labeled with the corresponding activity. The proposed decision-level fusion method was evaluated using five-fold cross-validation, with the data randomly split into training, validation, and testing datasets in a 3:1:1 ratio.

## 4. Results

### 4.1. Single-modal activity recognition results

The average testing accuracies from the five-fold cross-validation results are 85.6% and 85.4% for the activity recognition models trained from vision-based and acceleration-based algorithms, respectively. Table 3 and Table 4 present the confusion matrices with recall, precision, and F1 scores for each activity. For the vision-based model, activities such as “Hammering” (HR), “Idling” (ID), and “Drilling” (DR) achieved high performance, with precision, recall, and F1 scores exceeding 98%. However, activities like “Traveling” (TL), “Lifting Brick” (LB), and “Measuring Rebar” (MR) exhibited low performance, with F1 scores of 58.9%, 71.2%, and 84.2%, respectively. For the acceleration-based model, activities such as “Idling” (ID) and “Hammering” (HR) achieved high performance, with precision, recall, and F1 scores exceeding 89%. However, while “Measuring Rebar” (MR) exhibited lower performance (with all metrics below 80%), “Lifting Rebar” (LR) and “Tying Rebar” (TR) demonstrated relatively lower performance compared to other activities, with some metrics hovering around 80–83%.

### 4.2. Multi-modal activity recognition results

The current research employed and evaluated two proposed fusion approaches: the Dempster-Shafer (DS) method and the Category-wise Weighted Dempster-Shafer (CWDS). The fusion models were trained and evaluated using the same methods and data as used for the vision and acceleration models. The testing results of the two fusion models are presented in Table 5. As reported, the DS method significantly improved accuracy to 91.9%. Applying category-wise weighting and filtering out extreme predictions further enhanced accuracy, with results reaching up to 96.0%. The thresholds for filtering were determined by testing typical values, such as 0.1%, 1%, and 5%, and selecting the one that provided the best performance. Additionally, distinct types of weights based on accuracy-related metrics obtained from single-model training were tested to identify the most appropriate metric. In this test, the precision, recall, and F1 scores showed similar levels of improvement during the fusion procedure.

Table 6 presents the confusion matrix for the Category-wise Weighted Dempster-Shafer (CWDS) method, allowing for a detailed analysis of category-wise performance. The performance of all activity classifications reaches a high level. In particular, the activities DR, HR, ID, LB, LR, and TL have F1 scores exceeding 95%. The remaining activities did not reach F1 scores of 95% but are all over 90%, demonstrating balanced and robust classification capability. A comparison between multi-modal and single-modal approaches, along with a detailed analysis, will be included in the discussion.

## 5. Discussion

Experimental testing results, as indicated in Table 5, demonstrate that the proposed decision-level fusion method significantly increases activity recognition accuracy. Specifically, the CWDS method achieved 96.0% accuracy, which is up to 10% higher than single-modality models. This improvement validates the effectiveness of combining probabilistic outputs from single-modality models using the Dempster-Shafer approach, thereby enhancing overall classification capability. Additionally, applying weighting factors based on accuracy-related metrics of each model, such as the F1 score from the training process, effectively addressed the unequal credibility of the sensor modalities by reflecting each model’s category-specific strengths. This adjustment improved the overall accuracy to approximately 96.0%, representing about a 4% increase over the pure DST fusion method.

The main reason the fusion approach works is that different data modalities provide various features of the activity, exhibiting varying strengths and weaknesses. Moreover, in multi-class classification tasks such as those in this study, such strengths and weaknesses are category-specific. To better illustrate the comparison of the category-wise performance among vision-based, acceleration-based, and fusion models, the precision, recall, and F1 scores are presented in radar plots, as shown in Fig. 7. For instance, the acceleration-based model achieves higher performance in terms of precision, recall, and F1 score for the activity “Lifting Brick” (LB) than the vision-based model. In contrast, the vision-

**Table 3**
Confusion matrix of vision-based activity recognition result (1.0-s. window).

| True \ Predicted | DR | HR | ID | LB | LR | MR | TR | TL | Recall (%) |
|---|---|---|---|---|---|---|---|---|---|
| DR | 298 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100.0 |
| HR | 1 | 301 | 0 | 0 | 0 | 0 | 0 | 0 | 99.7 |
| ID | 3 | 0 | 299 | 0 | 0 | 0 | 0 | 0 | 99.0 |
| LB | 0 | 1 | 0 | 236 | 18 | 4 | 4 | 39 | 78.1 |
| LR | 0 | 0 | 0 | 13 | 271 | 10 | 2 | 6 | 89.7 |
| MR | 0 | 0 | 0 | 1 | 0 | 255 | 46 | 0 | 84.4 |
| TR | 0 | 0 | 0 | 1 | 0 | 31 | 270 | 0 | 89.4 |
| TL | 0 | 0 | 0 | 110 | 16 | 4 | 4 | 128 | 48.9 |
| Precision (%) | 98.7 | 99.7 | 100.0 | 65.4 | 88.9 | 83.9 | 82.8 | 74.0 | |
| F1 Score (%) | 99.3 | 99.7 | 99.5 | 71.2 | 89.3 | 84.2 | 86.0 | 58.9 | |

**Table 4**
Confusion matrix of acceleration-based action recognition result (1.0-s window).

| True \ Predicted | DR | HR | ID | LB | LR | MR | TR | TL | Recall (%) |
|---|---|---|---|---|---|---|---|---|---|
| DR | 271 | 4 | 9 | 0 | 0 | 12 | 2 | 0 | 90.9 |
| HR | 12 | 273 | 0 | 0 | 1 | 14 | 2 | 0 | 90.4 |
| ID | 15 | 0 | 278 | 1 | 1 | 1 | 2 | 4 | 92.1 |
| LB | 0 | 0 | 0 | 251 | 25 | 7 | 1 | 18 | 83.1 |
| LR | 0 | 0 | 2 | 12 | 251 | 23 | 7 | 7 | 83.1 |
| MR | 10 | 6 | 5 | 10 | 28 | 213 | 26 | 4 | 70.5 |
| TR | 4 | 5 | 2 | 7 | 10 | 17 | 251 | 6 | 83.1 |
| TL | 0 | 0 | 0 | 8 | 6 | 1 | 5 | 242 | 92.4 |
| Precision (%) | 86.9 | 94.8 | 93.9 | 86.9 | 78.0 | 74.0 | 84.8 | 86.1 | |
| F1 Score (%) | 88.9 | 92.5 | 93.0 | 84.9 | 80.4 | 72.2 | 83.9 | 89.1 | |

**Table 5**
Overall performance of construction activity recognition models utilizing various methodologies.

| Methodology | | | Average Testing Accuracy (%) |
|---|---|---|---|
| Single-modal Approaches | Vision-based | | 85.6 |
| | Acceleration-based | | 85.4 |
| Multi-modal Approaches (Decision-level Fusion) | Dempster-Shafer Theory (**DS**) | | 91.9 |
| | Category-wise Weighted Dempster-Shafer (**CWDS**) | Precision-based Weighting | 96.0 |
| | | Recall-based Weighting | 95.7 |
| | | F1 score-based Weighting | 95.1 |

based approach outperforms the acceleration-based method for the activity DR, with precision, recall, and F1 scores of 98.7%, 100.0%, and 99.3%, respectively, compared to 86.9%, 90.9%, and 88.9% for the acceleration-based model. The plots clearly show the improvement in classification performance from the CWDS method in most categories, such as TL, LB, and MR, but it does not always guarantee category-wise improvements. For example, the vision-based model achieves slightly higher precision for DR than both the acceleration-based and fusion models. This implies that the proposed fusion method not only compensates for the limitations of single-modal sources but may also

**Table 6**
Confusion matrix of decision-level fusion activity recognition using Category-wise Weighted Dempster-Shafer method.

| True \ Predicted | DR | HR | ID | LB | LR | MR | TR | TL | Recall (%) |
|---|---|---|---|---|---|---|---|---|---|
| DR | 298 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100.0 |
| HR | 1 | 299 | 0 | 0 | 0 | 1 | 1 | 0 | 99.0 |
| ID | 2 | 0 | 297 | 0 | 0 | 0 | 0 | 3 | 98.3 |
| LB | 0 | 0 | 0 | 285 | 8 | 0 | 0 | 9 | 94.4 |
| LR | 0 | 0 | 0 | 4 | 292 | 4 | 1 | 1 | 96.7 |
| MR | 3 | 0 | 0 | 0 | 0 | 282 | 17 | 0 | 93.4 |
| TR | 1 | 3 | 0 | 0 | 0 | 14 | 284 | 0 | 94.0 |
| TL | 0 | 0 | 0 | 8 | 3 | 0 | 0 | 251 | 95.8 |
| Precision (%) | 97.7 | 99.0 | 100.0 | 96.0 | 96.4 | 93.7 | 93.7 | 95.1 | |
| F1 Score (%) | 98.8 | 99.0 | 99.2 | 95.2 | 96.5 | 93.5 | 93.9 | 95.4 | |

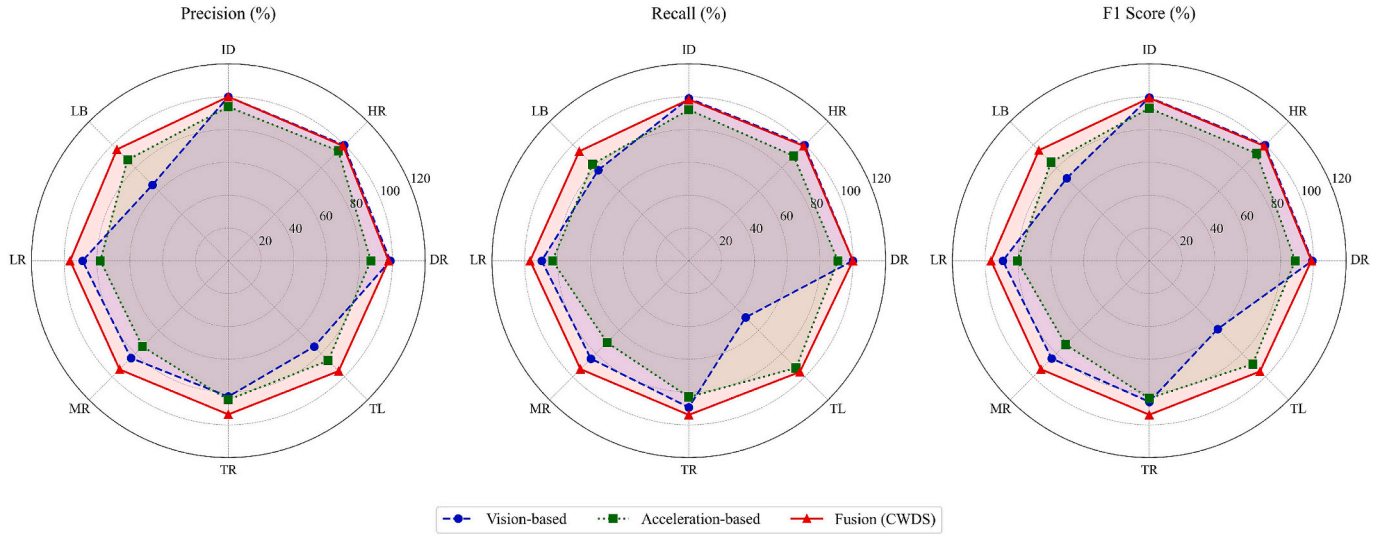



**Fig. 7.** Evaluation of single-modal and multi-modal (decision-level fusion) methods across activity categories.

introduce biases from other sources.

While the weighted method based on accuracy-related metrics obtained from each model's training has shown effectiveness in improving overall accuracy, different types of weighting factors (e.g., precision, recall, F1 scores) did not lead to significant differences. According to the CWDS equation, the probabilistic confidences of classes ($m_i$) from each model are combined with the weighting factors ($w_i$) that represent the category-wise strengths of each model by simply multiplying them. In general, the variability of $m_i$ (0.0–1.0) was much higher than that of $w_i$ (0.6–1.0). This may lead to less improvement when adopting weighting factors compared to applying probabilistic confidence.

Despite the advantages of the proposed decision-level fusion method for measuring the temporal aspects of manual handling activities, several limitations need to be addressed from both theoretical and practical perspectives. Although the proposed method improved accuracy, the roles and interaction effects of the two variables, $m_i$ and $w_i$, remain unclear. Further investigation is needed to identify how to optimize the combination of these two variables, incorporating advanced filtering and weighting techniques to mitigate biases and enhance accuracy. Additionally, as the experimental data were collected by repeating specific tasks independently, they may not fully reflect the potential data noise present in real-world conditions. For example, according to our previous study [26], most errors in acceleration-based activity recognition stem from activities that cannot be clearly defined due to the continuous nature of construction activities. Moreover, videos from construction sites are frequently occluded by other objects, leading to diminished performance in vision-based activity recognition. To address this issue, the proposed method will be further tested by collecting data on real construction sites.

## 6. Conclusion

This paper presents a methodology for automatically recognizing construction workers' activities by integrating video and acceleration data through a decision-level fusion approach. The proposed method leverages the complementary strengths of each modality, enabling more accurate activity recognition in construction settings. To develop and evaluate the model, training, validation, and test datasets comprising eight classes of construction-related activities were created in controlled lab environments. RGB videos were captured using smartphone cameras, and acceleration signals were recorded from wrist-worn sensors. Two independent deep-learning models were then used to perform preliminary activity recognition on each data modality. Category-specific weights were applied to the predictions to address differences in the reliability of each modality. These weights were estimated during the model training process and adjusted by filtering out extreme values. The weighted predictions were then combined using the Dempster-Shafer theory to determine the final activity classes.

The testing results demonstrated that the proposed method achieved up to 10% higher accuracy compared to single-modality approaches, as the Dempster-Shafer-based approach increased recognition accuracy from 85.6% and 85.4% to 91. %. The application of the weighted method further enhanced the overall performance to 96.0% by reflecting the category-specific strengths of each model. The proposed approach has also proven to improve performance in certain activity categories, such as lifting tasks. This improvement results from the ability to compensate for the weaknesses of each modality while preserving its unique strengths. Vision data provide detailed information about the lifting object. In contrast, acceleration data detect subtle motion responses caused by different types of lifting, whether involving the whole body or just the arm.

The proposed method offers several practical advantages for automatically measuring activities in productivity analysis. Single-modality approaches for activity recognition often fail due to the inherent limitations of individual sensor systems. In contrast, the proposed multimodal fusion approach enables activity recognition even when data from one sensor is missing, as the models using vision and acceleration data independently contribute to the final decision during the decision-level fusion stage. The independent structure also facilitates the updating of each single-modality model as needed. Furthermore, combining the outputs from vision- and acceleration-based recognition improves accuracy, allowing for more reliable activity analysis and better identification of potential issues related to low labor productivity. Although further improvements are necessary, such as parameter optimization, validation in field conditions, and testing on a larger and more diverse participant group, the proposed method is expected to serve as an effective tool for automatically collecting accurate field data during construction tasks using sensing technologies.

## CRediT authorship contribution statement

**Yue Gong:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **JoonOh Seo:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Conceptualization. **Kyung-Su Kang:** Validation, Methodology, Conceptualization. **Mengnan Shi:** Validation, Methodology.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## References

[1] M.A.R. Ahad, A.D. Antar, M. Ahmed, M.A.R. Ahad, A.D. Antar, M. Ahmed, Sensor-based human activity recognition: challenges ahead, IoT Sensor-Based Activity Recog.: Human Activity Recogn. (2021) 175–189, https://doi.org/10.1007/978-3-030-51379-5_10.

[2] R. Akhavian, A.H. Behzadan, Smartphone-based construction workers' activity recognition and classification, Autom. Constr. 71 (2016) 198–209, https://doi.org/10.1016/j.autcon.2016.08.015.

[3] I. Amundson, B. Kusy, P. Volgyesi, X. Koutsoukos, A. Ledeczi, Time synchronization in heterogeneous sensor networks, in: International Conference on Distributed Computing in Sensor Systems, Springer, 2008, pp. 17–31, https://doi.org/10.1007/1-84628-213-6_7.

[4] S.B. Ayed, H. Trichili, A.M. Alimi, Data fusion architectures: A survey and comparison, in: 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE, 2015, pp. 277–282, https://doi.org/10.1109/isda.2015.7489238.

[5] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495, https://doi.org/10.1109/TPAMI.2016.2644615.

[6] G. Badrinath, P. Gupta, Feature level fused ear biometric system, in: 2009 Seventh International Conference on Advances in Pattern Recognition, IEEE, 2009, pp. 197–200, https://doi.org/10.1109/icapr.2009.27.

[7] S.S. Bangaru, C. Wang, S.A. Busam, F.J. Aghazadeh, ANN-based automated scaffold builder activity recognition through wearable EMG and IMU sensors, Autom. Constr. 126 (2021) 103653, https://doi.org/10.1016/j.autcon.2021.103653.

[8] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, I. Rojas, Window size impact in human activity recognition, Sensors 14 (4) (2014) 6474–6499, https://doi.org/10.3390/s140406474.

[9] D.R. Beddiar, B. Nini, M. Sabokrou, A. Hadid, Vision-based human activity recognition: a survey, Multimed. Tools Appl. 79 (41) (2020) 30509–30555, https://doi.org/10.1007/s11042-020-09004-3.

[10] M. Beynon, D. Cosker, D. Marshall, An expert system for multi-criteria decision making using Dempster Shafer theory, Expert Syst. Appl. 20 (4) (2001) 357–367, https://doi.org/10.1016/S0957-4174(01)00020-3.

[11] J.S. Bohn, J. Teizer, Benefits and barriers of construction project monitoring using high-resolution automated cameras, J. Constr. Eng. Manag. 136 (6) (2010) 632–640, https://doi.org/10.1061/(Asce)Co.1943-7862.0000164.

[12] B. Buchholz, V. Paquet, H. Wellman, M. Forde, Quantification of ergonomic hazards for ironworkers performing concrete reinforcement tasks during heavy highway construction, AIHA J. 64 (2) (2003) 243–250, https://doi.org/10.1080/15428110308984814.

[13] F. Castanedo, A review of data fusion techniques, Sci. World J. 2013 (2013) 704504, https://doi.org/10.1155/2013/704504.

[14] A. Chehade, C. Song, K. Liu, A. Saxena, X. Zhang, A data-level fusion approach for degradation modeling and prognostic analysis under multiple failure modes, J. Qual. Technol. 50 (2) (2018) 150–165, https://doi.org/10.1080/00224065.2018.1436829.

[15] C. Chen, R. Jafari, N. Kehtarnavaz, Improving human action recognition using fusion of depth camera and inertial sensors, IEEE Transact. Human-Mach. Syst. 45 (1) (2014) 51–61, https://doi.org/10.1109/thms.2014.2362520.

[16] C. Chen, R. Jafari, N. Kehtarnavaz, A survey of depth and inertial sensor fusion for human action recognition, Multimed. Tools Appl. 76 (3) (2017) 4405–4425, https://doi.org/10.1007/s11042-015-3177-1.

[17] T. Cheng, J. Teizer, G.C. Migliaccio, U.C. Gatti, Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data, Autom. Constr. 29 (2013) 24–39, https://doi.org/10.1016/j.autcon.2012.08.003.

[18] B.R. Cobb, P.P. Shenoy, A comparison of Bayesian and belief function reasoning, Inf. Syst. Front. 5 (4) (2003) 345–358. Doi 10.1023/B:Isfi.0000005650.63806.03.

[19] A. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Stat. 38 (2) (1967) 325–339, https://doi.org/10.1214/aoms/1177698950.

[20] F. Dietrich, C. List, Majority voting on restricted domains, J. Econ. Theory 145 (2) (2010) 512–543, https://doi.org/10.1016/j.jet.2010.01.003.

[21] T.G. Dietterich, Machine Learning for Sequential Data: A Review, Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, 2002, pp. 15–30, https://doi.org/10.1007/3-540-70659-3_2.

[22] Y. Ding, X. Yao, S. Wang, X. Zhao, Structural damage assessment using improved Dempster-Shafer data fusion algorithm, Earthq. Eng. Eng. Vib. 18 (2) (2019) 395–408, https://doi.org/10.1007/s11803-019-0511-z.

[23] W. Elmenreich, An Introduction to Sensor Fusion vol. 502, Vienna University of Technology, Austria, 2002, pp. 1–28.

[24] V. Escorcia, M.A. Dávila, M. Golparvar-Fard, J.C. Niebles, Automated vision-based recognition of construction worker actions for building interior construction operations using RGBD cameras, in: Construction Research Congress 2012: Construction Challenges in a Flat World, 2012, pp. 879–888, https://doi.org/10.1061/9780784412329.089.

[25] J. Gong, C.H. Caldas, Computer vision-based video interpretation model for automated productivity analysis of construction operations, J. Comput. Civ. Eng. 24 (3) (2010) 252–263, https://doi.org/10.1061/(Asce)Cp.1943-5487.0000027.

[26] Y. Gong, K. Yang, J. Seo, J.G. Lee, Wearable acceleration-based action recognition for long-term and continuous activity analysis in construction site, J. Build. Eng. 52 (2022) 104448, https://doi.org/10.1016/j.jobe.2022.104448.

[27] J. Gordon, E.H. Shortliffe, The Dempster-Shafer theory of evidence, rule-based expert systems: the MYCIN experiments of the stanford heuristic programming project 3 (832–838) (1984) 3–4.

[28] M.C. Gouett, C.T. Haas, P.M. Goodrum, C.H. Caldas, Activity analysis for direct-work rate improvement in construction, J. Constr. Eng. Manag. 137 (12) (2011) 1117–1124, https://doi.org/10.1061/(Asce)Co.1943-7862.0000375.

[29] R. Gravina, P. Alinia, H. Ghasemzadeh, G. Fortino, Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges, Inform. Fusion 35 (2017) 68–80, https://doi.org/10.1016/j.inffus.2016.09.005.

[30] H. Gunes, M. Piccardi, Affect recognition from face and body: early fusion vs. late fusion, in: 2005 IEEE International Conference on Systems, Man and Cybernetics Vol. 4, IEEE, 2005, pp. 3437–3443, https://doi.org/10.1109/icsmc.2005.1571679.

[31] D.L. Hall, J. Llinas, An introduction to multisensor data fusion, Proc. IEEE 85 (1) (1997) 6–23, https://doi.org/10.1109/5.554205.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 770–778, https://doi.org/10.1109/cvpr.2016.90.

[33] Z. Hussain, M. Sheng, W.E. Zhang, Different approaches for human activity recognition: a survey, arXiv (2019), https://doi.org/10.48550/arXiv.1906.05074 preprint arXiv:1906.05074.

[34] E.L. Jacobsen, J. Teizer, S. Wandahl, Work estimation of construction workers for productivity monitoring using kinematic data and deep learning, Autom. Constr. 152 (2023) 104932, https://doi.org/10.1016/j.autcon.2023.104932.

[35] L. Joshua, K. Varghese, Accelerometer-based activity recognition in construction, J. Comput. Civ. Eng. 25 (5) (2011) 370–379, https://doi.org/10.1061/(Asce)Cp.1943-5487.0000097.

[36] L. Joshua, K. Varghese, Automated recognition of construction labour activity using accelerometers in field situations, Int. J. Product. Perform. Manag. 63 (7) (2014) 841–862, https://doi.org/10.1108/IJPPM-05-2013-0099.

[37] A. Khosrowpour, J.C. Niebles, M. Golparvar-Fard, Vision-based workface assessment using depth images for activity analysis of interior construction operations, Autom. Constr. 48 (2014) 74–87, https://doi.org/10.1016/j.autcon.2014.08.003.

[38] J. Kim, S. Chi, C.R. Ahn, Hybrid kinematic–visual sensing approach for activity recognition of construction equipment, J. Build. Eng. 44 (2021) 102709, https://doi.org/10.1016/j.jobe.2021.102709.

[39] K. Kim, Y.K. Cho, Effective inertial sensor quantity and locations on a body for deep learning-based worker's motion recognition, Autom. Constr. 113 (2020) 103126, https://doi.org/10.1016/j.autcon.2020.103126.

[40] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Proces. Syst. 25 (2012), https://doi.org/10.1145/3065386.

[41] D. Kuhse, N. Holscher, M. Gunzel, H. Teper, G. Von Der Bruggen, J.-J. Chen, C.-C. Lin, Sync or Sink? The robustness of sensor fusion against temporal misalignment, in: 2024 IEEE 30th Real-Time and Embedded Technology and Applications Symposium (RTAS), IEEE, 2024, pp. 122–134, https://doi.org/10.1109/rtas61025.2024.00018.

[42] S.C. Kulkarni, P.P. Rege, Pixel level fusion techniques for SAR and optical images: A review, Inform. Fusion 59 (2020) 13–29, https://doi.org/10.1016/j.inffus.2020.01.003.

[43] J. Li, Q. Miao, Z. Zou, H. Gao, L. Zhang, Z. Li, N. Wang, A review of computer vision-based monitoring approaches for construction Workers' work-related behaviors, IEEE Access 12 (2024) 7134–7155, https://doi.org/10.1109/Access.2024.3350773.

[44] Z. Li, D. Li, Action recognition of construction workers under occlusion, J. Build. Eng. 45 (2022) 103352, https://doi.org/10.1016/j.jobe.2021.103352.

[45] M. Liu, D. Hong, S. Han, S. Lee, Silhouette-based on-site human action recognition in single-view video, Construct. Res. Congr. 2016 (2016) 951–959, https://doi.org/10.1061/9780784479827.096.

[46] H. Luo, C. Xiong, W. Fang, P.E. Love, B. Zhang, X. Ouyang, Convolutional neural networks: computer vision-based workforce activity assessment in construction, Autom. Constr. 94 (2018) 282–289, https://doi.org/10.1016/j.autcon.2018.06.007.

[47] R.C. Luo, M.G. Kay, Multisensor integration and fusion: issues and approaches, in: Sensor Fusion Vol. 931, SPIE, 1988, pp. 42–49, https://doi.org/10.1117/12.946646.

[48] X. Luo, H. Li, D. Cao, F. Dai, J. Seo, S. Lee, Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks, J. Comput. Civ. Eng. 32 (3) (2018) 04018012, https://doi.org/10.1061/(Asce)Cp.1943-5487.0000756.

[49] R. Malhotra, Temporal considerations in sensor management, in: Proceedings of the IEEE 1995 National Aerospace and Electronics Conference. NAECON 1995 Vol. 1, IEEE, 1995, pp. 86–93, https://doi.org/10.1109/NAECON.1995.521917.

[50] A. Mannini, A.M. Sabatini, Machine learning methods for classifying human physical activity from on-body accelerometers, Sensors 10 (2) (2010) 1154–1175, https://doi.org/10.3390/s100201154.

[51] R. Navon, Automated project performance control of construction projects, Autom. Constr. 14 (4) (2005) 467–476, https://doi.org/10.1016/j.autcon.2004.09.006.

[52] L. Peng, L. Chen, M. Wu, G. Chen, Complex activity recognition using acceleration, vital sign, and location data, IEEE Trans. Mob. Comput. 18 (7) (2018) 1488–1498, https://doi.org/10.1109/TMC.2018.2863292.

[53] D. Roberts, W. Torres Calderon, S. Tang, M. Golparvar-Fard, Vision-based construction worker activity analysis informed by body posture, J. Comput. Civ. Eng. 34 (4) (2020) 04020017, https://doi.org/10.1061/(Asce)Cp.1943-5487.0000898.

[54] J. Ryu, J. Seo, H. Jebelli, S. Lee, Automated action recognition using an accelerometer-embedded wristband-type activity tracker, J. Constr. Eng. Manag. 145 (1) (2019) 04018114, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001579.

[55] S. Sathe, T.G. Papaioannou, H. Jeung, K. Aberer, A Survey of Model-Based Sensor Data Acquisition and Management, Managing and Mining Sensor Data, Springer, 2013, pp. 9–50, https://doi.org/10.1007/978-1-4614-6309-2_2.

[56] K. Sentz, S. Ferson, Combination of Evidence in Dempster-Shafer Theory, 2002, https://doi.org/10.2172/800792.

[57] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976, 069110042X.

[58] A. Sinha, H. Chen, D. Danu, T. Kirubarajan, M. Farooq, Estimation and decision fusion: A survey, Neurocomputing 71 (13–15) (2008) 2650–2656, https://doi.org/10.1016/j.neucom.2007.06.016.

[59] T. Slaton, C. Hernandez, R. Akhavian, Construction activity recognition with convolutional recurrent networks, Autom. Constr. 113 (2020) 103138, https://doi.org/10.1016/j.autcon.2020.103138.

[60] W. Sprague, E. Rezazadeh Azar, Integrating acceleration signal processing and image segmentation for condition assessment of asphalt roads, Can. J. Civ. Eng. 49 (6) (2022) 1095–1107, https://doi.org/10.1139/cjce-2021-0116.

[61] A.N. Steinberg, C.L. Bowman, Revisions to the JDL Data Fusion Model, Handbook of Multisensor Data Fusion, CRC press, 2017, pp. 65–88, 1315219484.

[62] X. Su, H. Tong, P. Ji, Activity recognition with smartphone sensors, Tsinghua Sci. Technol. 19 (3) (2014) 235–249, https://doi.org/10.1109/Tst.2014.6838194.

[63] B. Suo, L. Zhao, Y. Yan, A novel Dempster-Shafer theory-based approach with weighted average for failure mode and effects analysis under uncertainty, J. Loss Prev. Process Ind. 65 (2020) 104145, https://doi.org/10.1016/j.jlp.2020.104145.

[64] G.A.S. Surek, L.O. Seman, S.F. Stefenon, V.C. Mariani, L.D.S. Coelho, Video-based human activity recognition using deep learning approaches, Sensors 23 (14) (2023) 6384, https://doi.org/10.3390/s23146384.

[65] S.V.-T. Tran, T.L. Nguyen, H.-L. Chi, D. Lee, C. Park, Generative planning for construction safety surveillance camera installation in 4D BIM environment, Autom. Constr. 134 (2022) 104103, https://doi.org/10.1016/j.autcon.2021.104103.

[66] M. Tubaishat, S. Madria, Sensor networks: an overview, IEEE Potentials 22 (2) (2003) 20–23, https://doi.org/10.1109/MP.2003.1197877.

[67] P. Tzirakis, S. Zafeiriou, B. Schuller, Real-world Automatic Continuous Affect Recognition from Audiovisual Signals, Multimodal Behavior Analysis in the Wild, Elsevier, 2019, pp. 387–406, https://doi.org/10.1016/B978-0-12-814601-9.00028-6.

[68] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing, J. Li, A survey of multimodal sensor fusion for passive RF and EO information integration, IEEE Aerosp. Electron. Syst. Mag. 36 (7) (2021) 44–61, https://doi.org/10.1109/Maes.2020.3006410.

[69] H. Wei, R. Jafari, N. Kehtarnavaz, Fusion of video and inertial sensing for deep learning–based human action recognition, Sensors 19 (17) (2019) 3680, https://doi.org/10.3390/s19173680.

[70] H. Wu, M. Siegel, R. Stiefelhagen, J. Yang, Sensor fusion using Dempster-Shafer theory [for context-aware HCI], IMTC/2002, in: Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No. 00CH37276) Vol. 1, IEEE, 2002, pp. 7–12, https://doi.org/10.1109/IMTC.2002.1006807.

[71] T. Wu, L. Tang, S. Shao, X. Zhang, Y. Liu, Z. Zhou, X. Qi, Accurate structural displacement monitoring by data fusion of a consumer-grade camera and accelerometers, Eng. Struct. 262 (2022) 114303, https://doi.org/10.1016/j.engstruct.2022.114303.

[72] J.-Y. Yang, J.-S. Wang, Y.-P. Chen, Using acceleration measurements for activity recognition: an effective learning algorithm for constructing neural classifiers, Pattern Recogn. Lett. 29 (16) (2008) 2213–2220, https://doi.org/10.1016/j.patrec.2008.08.002.

[73] J. Yang, Enhancing action recognition of construction workers using data-driven scene parsing, J. Civ. Eng. Manag. 24 (7) (2018) 568–580, https://doi.org/10.3846/jcem.2018.6133.

[74] J. Yang, Z. Shi, Z. Wu, Vision-based action recognition of construction workers using dense trajectories, Adv. Eng. Inform. 30 (3) (2016) 327–336, https://doi.org/10.1016/j.aei.2016.04.009.

[75] Z. Yang, Y. Yuan, M. Zhang, X. Zhao, B. Tian, Assessment of construction workers' labor intensity based on wearable smartphone system, J. Constr. Eng. Manag. 145 (7) (2019) 04019039, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001666.

[76] P. Zappi, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, G. Troster, Activity recognition from on-body sensors by classifier fusion: sensor scalability and robustness, in: 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, IEEE, 2007, pp. 281–286, https://doi.org/10.1109/ISSNIP.2007.4496857.

[77] X. Zheng, M. Wang, J. Ordieres-Meré, Comparison of data preprocessing approaches for applying deep learning to human activity recognition in the context of industry 4.0, Sensors 18 (7) (2018) 2146, https://doi.org/10.3390/s18072146.

[78] D. Zhou, T. Wei, H. Zhang, S. Ma, F. Wei, An information fusion model based on Dempster–Shafer evidence theory for equipment diagnosis, ASCE-ASME J. Risk Uncertain. Eng. Syst. Part B: Mech. Eng. 4 (2) (2018) 021005, https://doi.org/10.1115/1.4037328.