Kalpa Publications in Computing

Volume 22, 2025, Pages 1058–1069

Proceedings of The Sixth International Conference on Civil and Building Engineering Informatics

Limited Data-oriented Worker Intention Recognition Method in Worker-Robot Collaboration for Construction

Zhaoxin Zhang [1*], Zaolin Pan [1†], Yantao Yu [1‡] and Liang Liu [2§]

[1] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR.

[2] Hong Kong INDICS International Information Technology Co., Ltd.

zhangzhaoxin@ust.hk, zpanaq@connect.ust.hk, ceyantao@ust.hk, liuliangyork@sina.com

Abstract

Timely and accurate identification of workers' intentions in construction scenarios is crucial for seamless worker-robot collaboration. However, limited worker behavior due to varying behavioral styles and difficulties in collecting worker action data limit the practical application of existing methods that rely heavily on extensive worker action data. This paper addresses the dynamic nature of construction environments by proposing a few-shot worker intention recognition method. The proposed approach constructs worker intention query features using randomly sampled frame combinations and then applies metric learning to develop a few-shot worker intention recognition model. To validate the effectiveness of this method, a worker scaffolding installation action video dataset was used for the experiments on worker intent recognition. Given five categories with five worker action samples, the method achieved an accuracy of 71% in recognizing workers' intentions. The results demonstrate that the proposed method can effectively learn and detect novel worker actions with a minimal number of classified action videos, thereby improving model performance while reducing the number of required training videos. This approach not only reduces the labor needed for data labeling but also enhances the practicality of worker-robot collaboration in construction scenarios.

[*] Manuscript writing, experimental design, data analysis, and result organization
[†] Overall guidance and revision of the manuscript
[‡] Code implementation and theoretical analysis
[§] Responsible for part of the experiments and data processing

I. Hkust, J. Cheng and Y. Yantao (eds.), ICCBEI 2025 (Kalpa Publications in Computing, vol. 22), pp. 1058–1069

1 Introduction

High-risk construction environments and frequent safety incidents pose significant obstacles to the high-quality and sustainable development of the construction industry (Z. Pan & Yu, 2024). These accidents subject workers to severe injuries and life-threatening hazards, while poor working conditions and safety risks exacerbate labor shortages in the sector. Therefore, there is an urgent need to improve safety protocols and address labor shortages.

Automation and robotics have emerged as critical solutions to these industry challenges (Baduge et al., 2022). However, the construction environment is highly unstructured and dynamic, with frequent changes even during task execution. Single-task robots lack the flexibility required to handle such uncertain workplaces, relying instead on human workers' decision-making and problem-solving abilities. To ensure the effective application of robots in this domain, Worker-Robot Collaboration (WRC) should be established—a novel collaborative framework that leverages the strengths of both workers and robots. WRC technology offers a promising solution to the pressing needs faced by the construction industry (Eaves et al., 2016). In the realm of construction, such collaboration can liberate construction workers from tasks characterized by repetitiveness and high physical demands, thereby enhancing construction safety and productivity, as well as mitigating labor shortages and the aging workforce crisis (Cai et al., 2023; Park et al., 2023). To facilitate seamless worker-robot collaboration, robots should possess the capability to intelligently perceive, comprehend, and adapt to the intentions of workers (Zhang et al., 2022), enabling assistance in task execution within dynamic construction environments.

However, most advanced technologies rely on conventional deep learning methods (S. Li et al., 2022), which require large amounts of training data with accurately labeled worker intention information to develop high-performance vision-based algorithms. Consequently, numerous construction scene images must be collected, and worker intention type labels must be annotated for each video (Tian et al., 2024). This manual process is time-consuming, costly, and labor-intensive, making the development of comprehensive and high-quality robotic perception databases a significant challenge (Teizer, 2015). The issue is further complicated by the need to recognize different worker intentions that frequently change with construction phases. Each time a new type of worker intention arises, the training database must be updated. Due to these factors, the performance of vision-based monitoring systems often deteriorates in practice.

To overcome the challenges of worker intention recognition with limited data, this study proposes a limited data-oriented worker intention recognition method in worker-robot collaboration for construction, which aims to minimize the required training data and reduce data labeling costs. The proposed method focuses on worker intention feature extraction and a feature learning strategy tailored to limited data scenarios. To validate the approach, experiments were conducted using a dataset of worker scaffold installation actions. Remarkably, even with only five worker action samples, the method achieved 71% accuracy in recognizing worker intentions.

The remainder of this paper is organized as follows: Section 2 provides a literature review of existing studies. Section 3 explains the worker intent recognition method under limited data in detail. Section 4 presents the recognition performance of the proposed method. Section 5 discusses the results, and Section 6 concludes the study.

2 Literature Review

2.1 Worker Intention Recognition

Human-robot collaboration (HRC) research explores the physical and cognitive interactions between humans and robots to accomplish shared tasks, such as closely collaborating with cooperative robots to safely and effectively complete various joint activities (Semeraro et al., 2022). In HRC, robots need to understand the current work status and worker intention (Y. Pan et al., 2023) to determine their own tasks and movements. Typical tasks include collaborative assembly (Cunha et al., 2020; Grigore et al., 2018)and object handover (Choi et al., 2018; Shukla et al., 2018). For instance, in (Cunha et al., 2020), the robot must comprehend the current steps of the assembly process and select the appropriate components for assembly. In (Grigore et al., 2018), the robot provides assistance based on predictions of human trajectories. For object handover, it is essential for the robot to understand human intention, as smooth collaboration requires the robot to accurately interpret the object desired by the worker (Shukla et al., 2018).

In the construction domain, HRC has gained increasing attention as a promising solution to free workers from repetitive and physically demanding tasks, thereby enhancing construction safety and productivity. It can potentially be applied to various construction activities, such as bricklaying (Y. Liu et al., 2021), object handover (Yu et al., 2023), and wooden component assembly (X. Wang et al., 2023). It is well known that the ability to understand ongoing work progress and interpret human intention is crucial for robots to adaptively collaborate and execute required tasks. Existing research has developed methods to guide robots in responding and executing tasks based on various sensors (e.g., tactile sensors (Yu et al., 2023), cameras (Wu et al., 2023), and electroencephalography (Y. Liu et al., 2021)). Specifically, human intention (e.g., target of interest), human posture, or the movement of specific body parts (e.g., hands) has been extensively studied through a range of sensory inputs, including visual (Z. Liu et al., 2019), accelerometric (H. Liu & Wang, 2017), muscular (W. Wang et al., 2022), and neural activities (Lyu et al., 2022).

2.2 Sample Learning Methods under Limited Data

Although traditional machine learning algorithms typically require large training datasets, humans demonstrate the ability to learn and recognize new objects from only a few examples. To address this gap, researchers have explored various few-shot learning algorithms to efficiently acquire construction-related knowledge. Few-shot learning has been applied to computer vision tasks, including few-shot image classification, few-shot object detection, and few-shot video action recognition (Song et al., 2023). Few-shot video action recognition aims to achieve video classification capabilities by providing only a limited number of video sequences. Existing studies have effectively explored transfer learning (Hu et al., 2022; Khacef et al., 2020; H. Wang et al., 2021), meta-learning (Finn et al., 2017; Patacchiola et al., 2020; J. Xu et al., 2020; Zhong et al., 2020), and metric learning (Cao et al., 2020; X. Li et al., 2023; Perrett et al., 2021; van der Spoel et al., 2015)for few-shot data. Transfer learning algorithms operate on the assumption that knowledge obtained from data in the source domain can be transferred to the target domain through model fine-tuning. Meta-learning algorithms aim to learn generalizable knowledge by training models on several different classification tasks. Metric learning algorithms focus on creating feature embeddings and effective distance measures.

Few-shot action recognition has been validated on public computer vision datasets; however, there is a research gap in conducting few-shot action recognition on challenging construction sites. These sites are characterized by high dynamism and privacy sensitivity, making it difficult to collect large datasets of labeled video sequences. Notably, labeling data related to worker action intentions requires substantial expertise and additional time costs. Due to factors such as complex background variations, similarities between different actions, and occlusion of objects on-site, recognizing worker intent remains challenging (Zhang et al., 2022). Worker intent may change depending on the construction environment and the assembly components involved. Recent studies have applied few-shot learning techniques to construction site data. For example, Cui et al. proposed using few-shot classification and contrastive learning to identify facade defects (Cui et al., 2022). Xu et al. developed a meta-learning-based few-shot classifier for recognizing structural damage from images (Y. Xu et al., 2021). Kim et al. applied few-shot object detection to learn and monitor emerging targets, such as excavators on construction sites (Kim & Chi, 2021). Liang et al. proposed a CLIP-based few-shot learning algorithm for temporary object recognition on construction sites (Liang et al., 2024). Wang et al. proposes a deep learning-based method to detect fall-related site objects and their associated attributes to better capture site conditions for supporting field compliance checking (X. Wang & El-Gohary, 2024).

3 Methodology

The proposed method aims to design a few-shot model for worker intention recognition that can learn and classify with only a limited dataset of worker intention.

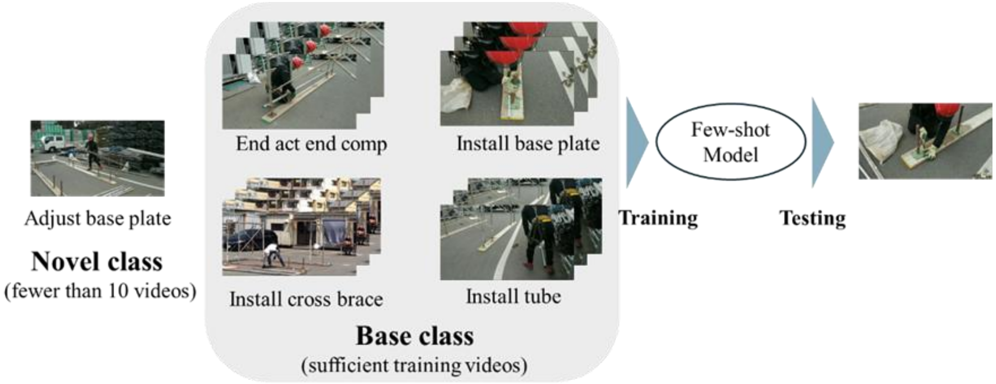

Figure 1: Concept of few-shot learning

As illustrated in Figure 1, the proposed method aims to generate a model that can learn and classify worker intention based on a limited amount of video data. First, we define the variables used in this study. The objective of this paper is to accurately classify worker intentions that the model has not previously encountered. The small amount of labeled data during the model training phase is referred to as the 'support set.' We categorize each class in the few-shot scenario into k intention categories and n video data, commonly known as a k-way n-shot detection scenario. Let $Q = \{q_1, q_2, \dots, q_F\}$ denote a single query video with F frames. The goal is to classify Q into one of the classes$c \in C$. For class c, its support set S^c contains K videos, with the k^{th} video defined as $S_k^c = \{s_{k1}^c, s_{k2}^c, \dots, s_{kF}^c\}$.

3.1 Model Design

This process designs a few-shot learning model capable of acquiring meta-knowledge from training videos labeled with only base classes (i.e., how to learn worker intention from video data) and leveraging this knowledge to learn new types of worker intention. The proposed method consists of video feature extraction and prediction modules.

First, the video feature extraction module learns to extract generalized visual features that can represent various worker intentions. These generalized features, referred to as meta-features, delineate and elucidate the overall characteristics of worker intention types within the dataset, as shown in Figure 2.

In this study, the module consists of two components. In the first step, convolutional neural networks are employed to extract features from individual frames of worker operations. For this single-frame image feature extraction, ResNet-34 is utilized for image feature computation.

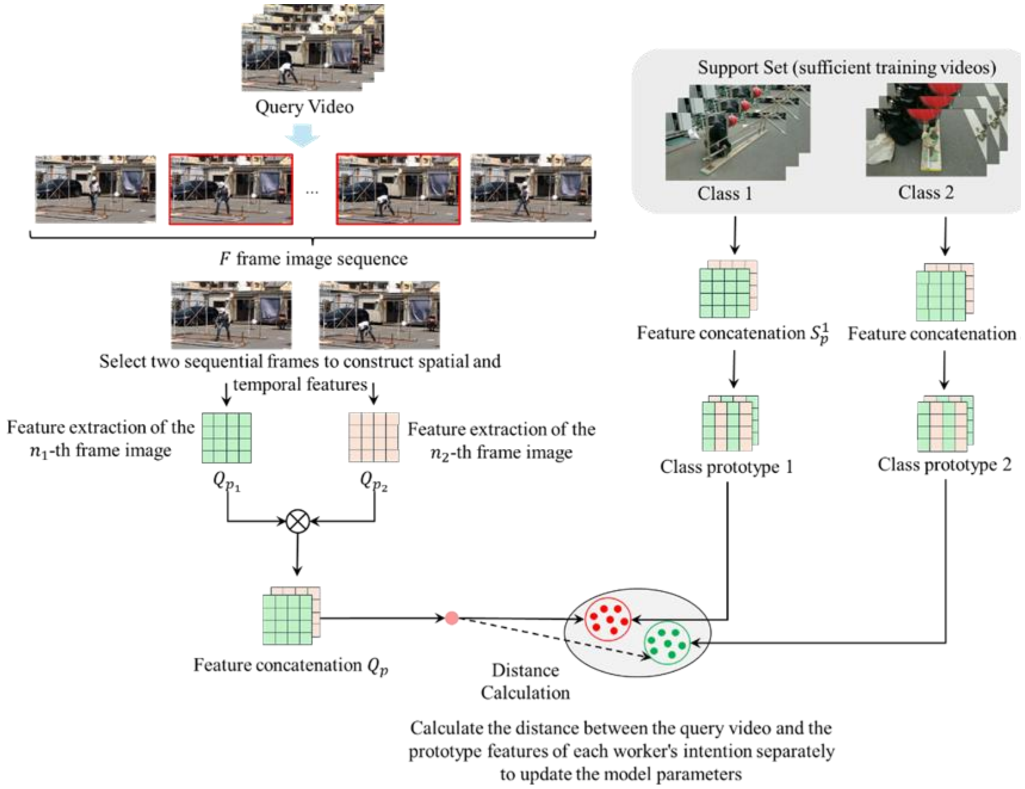
Query Video
Support Set (sufficient training videos)
Class 1
Class 2
...
F frame image sequence
Feature concatenation S_p^1
Feature concatenation
Select two sequential frames to construct spatial and temporal features
Feature extraction of the n_1-th frame image
Feature extraction of the n_2-th frame image
Q_{p_1}
Q_{p_2}
Class prototype 1
Class prototype 2
Feature concatenation Q_p
Distance Calculation
Calculate the distance between the query video and the prototype features of each worker's intention separately to update the model parameters

Figure 2: The architecture of the proposed few-shot worker intention recognition model.

In the second step, worker intention may exhibit different feature representations at different moments, making it challenging to capture complex features using only single frames. Furthermore, due to the complexities and dynamic nature of construction scenes, such as worker occlusion, worker intention features may not be fully captured by the robot. This leads to issues of discontinuity in the image sequences of worker intention and variations in action speed from the robot's perspective. To address the inter-frame sequence feature extraction, this study adopts a method of randomly combining two frames from the worker intention video to extract inter-frame sequence features, thereby mitigating the impact of discontinuous video sequences on the model.

Regarding the query video Q, we randomly slice the video frame sequence and set $p = (p_1, p_2)$. The video feature extraction calculation formula is as follows:

$$Q_p = \left[\Phi\left(q_{p_1}\right), \Phi\left(q_{p_2}\right)\right] \tag{1}$$

where $\Phi()$ represents the convolutional neural network.

For the support video S_{km}^c, the feature representation is given by:

$$S_{km}^{c} = \left[\Phi\left(s_{km_1}^{c}\right), \Phi\left(s_{km_2}^{c}\right)\right] \tag{2}$$

where $m = (m_1, m_2)$.

3.2 Similarity Calculation

Similarity calculation is a crucial algorithm in few-shot models based on metric learning. This study aims to learn an appropriate similarity measure that enables better differentiation between worker intention types in the original task. Such methods do not rely on extensive parameter updates but instead make predictions through a fixed metric approach. They guide the Few-Shot model in learning how to (1) extract worker intention image features and inter-frame features from input videos using only labeled training data from the support set, and (2) compute their category prototypes when provided with some support images, subsequently determining the current worker intention type by calculating the distance between the query video and the category prototypes.

The metric learning-based few-shot worker intention recognition model plays a significant role in learning how to obtain an appropriate metric space from video frame images. Therefore, to enhance the model's feature generalization performance, during the training phase, the model utilizes sufficiently labeled training data to simulate few-shot detection scenarios. The model is then trained based on the provided few-shot data. In this study, the worker intention category prototype S_c is obtained using the temporal CrossTransformer computational method (Perrett et al., 2021). Let Q_i denote the query worker intention features. The distance for each query is taken negatively as the loss function for model parameter updates.

We utilize the Euclidean metric to calculate the distance between support features S_c and query features Q_i :

$$d = \|Q_i - S_c\|^2 \tag{3}$$

Based on the distance d, the model's loss function in the C-way K-shot task can be computed as:

$$\hat{d}_i^c = \frac{e^{-d_i^c}}{\sum_{c=1}^{C} e^{-d_i^c}} \tag{4}$$

During the training phase, we minimize the loss function to update the parameters of the proposed network, repeating this process across all randomly sampled tasks.

$$L = -\frac{1}{C}\sum_{i=1}^{C} \log\left(\hat{d}_i^c\right) \tag{5}$$

During the testing phase, we first need to configure different video categories, and a limited amount of data based on the input, which constitutes the k-way n-shot detection scenario. Using the few-shot model parameters obtained during the training phase, we derive the category prototypes for k worker intentions. The class with the minimum distance is then selected as the identified worker intention type. The worker intention recognition during the testing phase can be expressed as $arg\,\underset{c}{min}\,\hat{d}_i^c$. Unlike the training phase, new worker intention categories do not appear during training; they represent entirely new worker intention categories. In the testing phase, the worker intention category prototypes for the new classes are calculated solely using the trained model weights.

4 Experimental Results and Analysis

To validate the proposed method, two distinct experiments were conducted: one involving few-shot learning and the other involving traditional supervised learning. Supervised learning is a conventional method for recognizing worker intent, utilizing numerous support images for model training, while few-shot learning focuses on acquiring general knowledge of worker intent by learning a feature prototype from the provided support video set. This study employed SLOWFAST (Feichtenhofer et al., 2019) and X3D (Feichtenhofer, 2020) architectures as baseline models for traditional supervised learning. Comparing with the architectures of conventional supervised learning allows for a more comprehensive assessment of the impact of few-shot learning. Consequently, the authors can confirm the applicability and practicality of few-shot learning and objectively evaluate its beneficial effects.

For the experiments and result analysis, the research team utilized a dataset from the literature, which consists of scenes collected from YouTube, involving a worker operating on the same type of scaffolding with varying backgrounds and camera angles. In the authors' experiments, a total of 22 worker operation video data were randomly divided into training and testing datasets. Using this data, the research team trained and tested the model in different k-way n-shot scenarios, where the number of categories k was set to 2, 3, 4, 5, 10, and 20, and the number of support images n (given for each new category) was set to 1, 2, 3, 5, and 10. To minimize the loss function, stochastic gradient descent with a batch size of 8 was employed to train the few-shot worker intent recognition model, with a total of 30,000 iterations and a learning rate of 0.001. The performance of the trained model was evaluated using classification accuracy, which is one of the most widely used metrics for vision-based action classification networks.

4.1 Performance of the Proposed Method

Table 1 presents the experimental results based on different few-shot detection scenarios (k=2,3,4,5,10,20 and n=1,2,3,5,10). The developed model achieved a worker intent recognition accuracy of 71% in the 5-way 5-shot scenario. When 10 labeled videos of new worker intent classes were provided, the accuracy for these new classes reached 80.6%. Furthermore, as the amount of video data for worker intent recognition increased, the accuracy for recognizing new classes progressively improved.

	1-shot	2-shot	3-shot	5-shot	10-shot
2-way	71.8	77.9	81.2	87.8	91.3
3-way	58.8	69.2	72.9	80.9	86.6
4-way	53.2	62.1	67.8	75.5	82.5
5-way	47.7	58.1	64	71	80.6
10-way	36.2	45.7	51.8	60.7	71.3
20-way	28.5	37.8	44.8	53.2	64.6

Table 1: Experimental Results of the Proposed Method

4.2 Performance with and Without Few-Shot Learning

The authors also observed a significant positive impact of the proposed few-shot learning approach from the experiments. As shown in Figure 3, under the 10-way conditions, the recognition performance for new classes using few-shot learning consistently outperformed existing supervised learning methods (i.e., SLOWFAST and X3D). These results align with the established knowledge in computer vision and deep learning, which suggests that complex models cannot be effectively trained when training data is limited. The performance gap between few-shot learning and the best-performing supervised learning model (i.e., X3D) averaged approximately 35.0%, with these differences increasing as the values of k and n rise. This may be attributed to the requirement of supervised learning methods for a substantial amount of labeled training data across all classes, while learning new classes with limited video data poses challenges. In contrast, few-shot learning can leverage a few provided samples to "learn how to recognize new classes." These findings indicate that few-shot learning not only reduces the amount of training data necessary for generating deep learning models but also enables rapid learning for detecting new categories, even with minimal training data. Given these advantages, it may be applicable in real construction environments where numerous novel and foundational objects coexist.

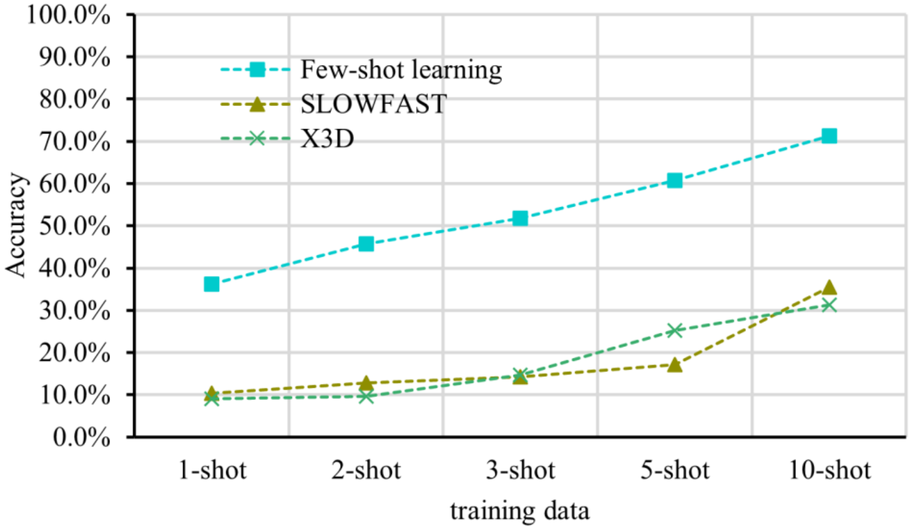

Figure 3: Comparative analysis for the accuracy of few-shot and supervised learning

5 Results and Discussion

The research findings indicate that the proposed method can learn meta-knowledge (i.e., how to learn target features) from a limited amount of labeled training data and utilize this knowledge to detect new worker intents when provided with some support videos. The model performed well despite variations in camera positions and viewpoints that presented different worker postures and visual features (e.g., color, shape, size).

Furthermore, this study shows that a promising deep learning model can be constructed using only a limited amount of training data. These findings could positively impact the implementation of vision-based monitoring in real-world construction sites. Specifically, the proposed few-shot learning method can quickly adapt to new human-robot collaboration needs arising at different construction stages, enabling the monitoring of new worker intentions without requiring extensive training data. This capability allows administrators to reduce the time and effort needed for additional data collection and annotation, which can be highly labor-intensive.

The technical advantages of this approach will be particularly useful during the phased development and updating of training databases. On-site managers can implement vision-based monitoring more effectively, as the development and updating of datasets represent one of the critical barriers to the use of visual systems in their projects. Additionally, the theoretical insights from this study can provide valuable guidance and future directions for researchers in the field of vision-based human-robot collaboration.

6 Conclusions

This study proposes a limited data-oriented worker intention recognition method in worker-robot collaboration for construction, aimed at minimizing the amount of training data and the cost of data labeling. The proposed method includes worker intention feature extraction and an intention classification strategy for limited data. As observed in the experiments, the few-shot model can learn meta-knowledge from a limited amount of labeled training data, and meta-learning successfully transfers this knowledge to detect new categories when provided with several video samples. In the 10-way 10-shot scenario, the accuracy for detecting new categories reached 71.3%, whereas the performance of supervised learning was limited to 35.5%. This indicates that the proposed method can train a more robust worker intention recognition model with the same amount of training data compared to existing methods.

The benefits of the proposed method contribute to the following advancements: First, this research introduces a novel technical framework that significantly reduces the number of required training images while maintaining performance in vision-based worker intention recognition. Second, the framework saves time and costs associated with data labeling, enhancing the practical acceptability of vision systems on construction sites. The authors customized few-shot learning for construction environments and provided quantitative results demonstrating its technical feasibility. Finally, the findings of this study lay the groundwork for future research in vision-based construction human-robot collaboration, as well as for other research areas, such as model-based construction automation monitoring.

7 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.

8 Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant number NSFC22EG19]; The Hong Kong University of Science and Technology 30 for 30 Research Fund [grant number 3030007]; the Research Grants Council (Hong Kong) [grant number 26208323]; University Grants Committee of The Government of the Hong Kong Special Administrative Region Collaborative Research Fund [grant number C6044-23GF].

References

Baduge, S. K., Thilakarathna, S., Perera, J. S., Arashpour, M., Sharafi, P., Teodosio, B., Shringi, A., & Mendis, P. (2022). Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. In *Automation in Construction* (Vol. 141). https://doi.org/10.1016/j.autcon.2022.104440

Cai, J., Du, A., Liang, X., & Li, S. (2023). Prediction-Based Path Planning for Safe and Efficient Human–Robot Collaboration in Construction via Deep Reinforcement Learning. *Journal of Computing in Civil Engineering*, *37*(1). https://doi.org/10.1061/(asce)cp.1943-5487.0001056

Cao, K., Ji, J., Cao, Z., Chang, C. Y., & Niebles, J. C. (2020). Few-shot video classification via temporal alignment. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR42600.2020.01063

Choi, S., Lee, K., Park, H. A., & Oh, S. (2018). A Nonparametric Motion Flow Model for Human Robot Cooperation. *Proceedings - IEEE International Conference on Robotics and Automation*. https://doi.org/10.1109/ICRA.2018.8463201

Cui, Z., Wang, Q., Guo, J., & Lu, N. (2022). Few-shot classification of façade defects based on extensible classifier and contrastive learning. *Automation in Construction*, *141*. https://doi.org/10.1016/j.autcon.2022.104381

Cunha, A., Ferreira, F., Sousa, E., Louro, L., Vicente, P., Monteiro, S., Erlhagen, W., & Bicho, E. (2020). Towards collaborative robots as intelligent co-workers in human-robot joint tasks: What to do and who does it? *52nd International Symposium on Robotics, ISR 2020*.

Eaves, S., Gyi, D. E., & Gibb, A. G. F. (2016). Building healthy construction workers: Their views on health, wellbeing and better workplace design. *Applied Ergonomics*, *54*. https://doi.org/10.1016/j.apergo.2015.11.004

Feichtenhofer, C. (2020). X3D: Expanding Architectures for Efficient Video Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR42600.2020.00028

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-October*. https://doi.org/10.1109/ICCV.2019.00630

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *34th International Conference on Machine Learning, ICML 2017*, *3*.

Grigore, E. C., Roncone, A., Mangin, O., & Scassellati, B. (2018). Preference-Based Assistance Prediction for Human-Robot Collaboration Tasks. *IEEE International Conference on Intelligent Robots and Systems*. https://doi.org/10.1109/IROS.2018.8593716

Hu, S. X., Li, D., Stuhmer, J., Kim, M., & Hospedales, T. M. (2022). Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*. https://doi.org/10.1109/CVPR52688.2022.00886

Khacef, L., Gripon, V., & Miramond, B. (2020). GPU-Based Self-Organizing Maps for Post-labeled Few-Shot Unsupervised Learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12533 LNCS*. https://doi.org/10.1007/978-3-030-63833-7_34

Kim, J., & Chi, S. (2021). A few-shot learning approach for database-free vision-based monitoring on construction sites. *Automation in Construction*, *124*. https://doi.org/10.1016/j.autcon.2021.103566

Li, S., Zheng, P., Fan, J., & Wang, L. (2022). Toward Proactive Human-Robot Collaborative Assembly: A Multimodal Transfer-Learning-Enabled Action Prediction Approach. *IEEE Transactions on Industrial Electronics*, *69*(8). https://doi.org/10.1109/TIE.2021.3105977

Li, X., Yang, X., Ma, Z., & Xue, J. H. (2023). Deep metric learning for few-shot image classification: A Review of recent developments. In *Pattern Recognition* (Vol. 138). https://doi.org/10.1016/j.patcog.2023.109381

Liang, Y., Vadakkepat, P., Chua, D. K. H., Wang, S., Li, Z., & Zhang, S. (2024). Recognizing temporary construction site objects using CLIP-based few-shot learning and multimodal prototypes. *Automation in Construction*, *165*, 105542. https://doi.org/https://doi.org/10.1016/j.autcon.2024.105542

Liu, H., & Wang, L. (2017). Human motion prediction for human-robot collaboration. *Journal of Manufacturing Systems*, *44*, 287–294. https://doi.org/https://doi.org/10.1016/j.jmsy.2017.04.009

Liu, Y., Habibnezhad, M., & Jebelli, H. (2021). Brainwave-driven human-robot collaboration in construction. *Automation in Construction*, *124*. https://doi.org/10.1016/j.autcon.2021.103556

Liu, Z., Liu, Q., Xu, W., Liu, Z., Zhou, Z., & Chen, J. (2019). Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing. *Procedia CIRP*, *83*. https://doi.org/10.1016/j.procir.2019.04.080

Lyu, J., Maýe, A., Görner, M., Ruppel, P., Engel, A. K., & Zhang, J. (2022). Coordinating human-robot collaboration by EEG-based human intention prediction and vigilance control. *Frontiers in Neurorobotics*, *16*. https://doi.org/10.3389/fnbot.2022.1068274

Pan, Y., Chen, C., Zhao, Z., Hu, T., & Zhang, J. (2023). Robot teaching system based on hand-robot contact state detection and motion intention recognition. *Robotics and Computer-Integrated Manufacturing*, *81*. https://doi.org/10.1016/j.rcim.2022.102492

Pan, Z., & Yu, Y. (2024). Learning multi-granular worker intentions from incomplete visual observations for worker-robot collaboration in construction. *Automation in Construction*, *158*. https://doi.org/10.1016/j.autcon.2023.105184

Park, S., Yu, H., Menassa, C. C., & Kamat, V. R. (2023). A Comprehensive Evaluation of Factors Influencing Acceptance of Robotic Assistants in Field Construction Work. *Journal of Management in Engineering*, *39*(3). https://doi.org/10.1061/jmenea.meeng-5227

Patacchiola, M., Turner, J., Crowley, E. J., O'Boyle, M., & Storkey, A. (2020). Bayesian meta-learning for the few-shot setting via deep kernels. *Advances in Neural Information Processing Systems*, *2020-December*.

Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., & Damen, D. (2021). Temporal-Relational CrossTransformers for Few-Shot Action Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR46437.2021.00054

Semeraro, F., Griffiths, A., & Cangelosi, A. (2022). Human–robot collaboration and machine learning: A systematic review of recent research. In *Robotics and Computer-Integrated Manufacturing* (Vol. 79). https://doi.org/10.1016/j.rcim.2022.102432

Shukla, D., Erkent, O., & Piater, J. (2018). Learning semantics of gestural instructions for human-robot collaboration. *Frontiers in Neurorobotics*, *12*(MAR). https://doi.org/10.3389/fnbot.2018.00007

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities. *ACM Computing Surveys*, *55*(13s). https://doi.org/10.1145/3582688

Teizer, J. (2015). Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. *Advanced Engineering Informatics*, *29*(2). https://doi.org/10.1016/j.aei.2015.03.006

Tian, Y., Chen, J., Kim, J. I., & Kim, J. (2024). Lightweight deep learning framework for recognizing construction workers' activities based on simplified node combinations. *Automation in Construction*, *158*. https://doi.org/10.1016/j.autcon.2023.105236

van der Spoel, E., Rozing, M. P., Houwing-Duistermaat, J. J., Eline Slagboom, P., Beekman, M., de Craen, A. J. M., Westendorp, R. G. J., & van Heemst, D. (2015). Siamese Neural Networks for One-Shot Image Recognition. *ICML - Deep Learning Workshop*, *7*(11).

Wang, H., Zhao, H., & Li, B. (2021). Bridging Multi-Task Learning and Meta-Learning: Towards Efficient Training and Effective Adaptation. *Proceedings of Machine Learning Research*, *139*.

Wang, W., Li, R., Chen, Y., Sun, Y., & Jia, Y. (2022). Predicting Human Intentions in Human-Robot Hand-Over Tasks Through Multimodal Learning. *IEEE Transactions on Automation Science and Engineering*, *19*(3). https://doi.org/10.1109/TASE.2021.3074873

Wang, X., & El-Gohary, N. (2024). Few-shot object detection and attribute recognition from construction site images for improved field compliance. *Automation in Construction, 167*, 105539. https://doi.org/https://doi.org/10.1016/j.autcon.2024.105539

Wang, X., Wang, S., Menassa, C. C., Kamat, V. R., & McGee, W. (2023). Automatic high-level motion sequencing methods for enabling multi-tasking construction robots. *Automation in Construction*, *155*. https://doi.org/10.1016/j.autcon.2023.105071

Wu, H., Li, H., Chi, H. L., Peng, Z., Chang, S., & Wu, Y. (2023). Thermal image-based hand gesture recognition for worker-robot collaboration in the construction industry: A feasible study. *Advanced Engineering Informatics*, *56*. https://doi.org/10.1016/j.aei.2023.101939

Xu, J., Ton, J. F., Kim, H., Kosiorek, A. R., & Teh, Y. W. (2020). MetaFun: Meta-learning with iterative functional updates. *37th International Conference on Machine Learning, ICML 2020, PartF168147-14*.

Xu, Y., Bao, Y., Zhang, Y., & Li, H. (2021). Attribute-based structural damage identification by few-shot meta learning with inter-class knowledge transfer. *Structural Health Monitoring*, *20*(4). https://doi.org/10.1177/1475921720921135

Yu, H., Kamat, V. R., Menassa, C. C., McGee, W., Guo, Y., & Lee, H. (2023). Mutual physical state-aware object handover in full-contact collaborative human-robot construction work. *Automation in Construction*, *150*. https://doi.org/10.1016/j.autcon.2023.104829

Zhang, Y., Ding, K., Hui, J., Lv, J., Zhou, X., & Zheng, P. (2022). Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly. *Advanced Engineering Informatics*, *54*. https://doi.org/10.1016/j.aei.2022.101792

Zhong, X., Gu, C., Huang, W., Li, L., Chen, S., & Lin, C. W. (2020). Complementing representation deficiency in few-shot image classification: A meta-learning approach. *Proceedings - International Conference on Pattern Recognition*. https://doi.org/10.1109/ICPR48806.2021.9412416