Learning multi-granularity task primitives from construction videos for human-robot collaboration

Zaolin PAN, Yantao YU
The Hong Kong University of Science and Technology, Hong Kong, China
zpanaq@connect.ust.hk, ceyantao@ust.hk

Abstract. Human-robot collaboration (HRC) is an emerging solution for productivity and safety concerns in the construction sector. The seamless HRC requires robots with task awareness and explainable perception processes. However, the implicit, dynamic construction task flow and noisy visual data captured in the field pose non-trivial challenges. To address these challenges, a vision-based multi-granularity task primitive learning method is proposed. Specifically, this study seeks to enhance the mutual understanding between workers and robots by determining which granularity level is best for the tasks' understanding and whether the robot learns useful visual cues to solve the task. Results show that the intermediate level has the best compromise between classification performance and task knowledge and that the model learns both useful and useless cues to recognise task primitives. These results will increase our understanding of multi-granular worker behavior and robot perception processes. The outcomes will improve the smoothness of HRC teams.

1. Introduction

Automation and robotics have become mainstream solutions to productivity and safety challenges facing the construction industry (Brosque *et al.*, 2020). It has been reported that more than 50 single-task construction robots have been developed to automate repetitive, physically demanding, and hazardous construction tasks (Bock and Linner, 2016). Although advances in sensing, manipulation, and computing enable the application of robotics in construction (Brosque *et al.*, 2020), worker assistance and supervision are essential. Robots alone are not capable of accomplishing construction tasks, primarily due to the diversity and dynamic nature of the tasks, which necessitate the timely decision-making and flexible task-handling skills of workers (Luo *et al.*, 2020; Liu and Jebelli, 2022). To enable effective robot application in the field, a new paradigm, i.e., Human-Robot Collaboration (HRC), combining the dexterity and knowledge of humans with the strength and speed of industrial robots, should be established.

To enable seamless HRC, robots need to be task-aware. Task knowledge is a critical prerequisite for robots to recognise worker intentions and provide tailored assistance (Grigore *et al.*, 2018). However, unlike other industries, the construction industry has few standard task flows. Workers can adapt their execution processes based on environmental conditions (Wu *et al.*, 2022). In order to handle the complexity of construction tasks, we formalise them as being composed of basic units of a task called task primitives. Depending on the level of abstraction, task primitives can be divided into different granularities, such as high-level subtasks and low-level actions. To understand the dynamic construction task, the robot should learn at different levels. Learning at a high level informs robots about the task's progression, and learning at a low level tells how workers interact with the environment. However, finding a clear and useful granularity level can be challenging. Although a higher level can produce good classification performance, it may conceal multiple subclasses. If we choose a lower level, recognition performance suffers due to the diverse worker actions. Hence, discovering an appropriate granularity of task primitives is essential for gaining a better understanding of the task at hand.

To foster human understanding of robots, on the other hand, it is crucial to understand what a robot learns from the visual world. The visual data captured in the field will inevitably contain noise and interference because of the complex and dynamic environment. If robots make decisions based on this irrelevant information, it can compromise the versatility of robots, undermine worker trust in robots, and even endanger worker safety. One viable option to address this problem is to provide visual explanations for the perception processes of robots (Anjomshoae *et al.*, 2019). Visual explanations of what robots have learned from the visual world to solve the task allow us to assess the reliability and reasonableness of robot decisions. This helps increase trust, smoothness, and productivity in a HRC team. Therefore, we expect that the task primitive learning approach can generate visual explanations.

Existing studies on learning task primitives can be divided into rule-based and data-based methods. Rule-based methods generally recognise task primitives through manually defined rules (Martinez *et al.*, 2021). These methods have good performance and interpretability in recognising task primitives that involve large equipment (Gong and Caldas, 2010). However, defining rules for learning task primitives related to workers' activities can be challenging due to the diversity of workers' actions (Luo *et al.*, 2020). Data-based methods address this challenge via a data-driven strategy. These methods leverage deep learning models to directly learn task primitives from visual data, showing promising performance in low- and high-level task primitive learning (Luo *et al.*, 2020). Nonetheless, existing data-based methods have the following limitations: (1) they focus on a single-granular and easily labelled task primitive, without access to holistic task information to determine which granularity level would be more appropriate for the task; and (2) they pay more attention to improving model performance on datasets while ignoring the interpretation of predictions, and thus, what the model learns from the visual world to accomplish the task remains unknown.

To address these limitations, we aim to learn multi-granularity task primitives from construction videos for HRC. The objectives are: (1) to design a multi-granularity task representation to model the construction task structure; (2) to develop a multi-granularity task primitive learning model for determining which granularity of task primitives is most appropriate for the task at hand; (3) to visualise the learned visual cues to find out whether the model learned useful visual cues or dataset biases to solve the task; and (4) to test and evaluate the framework on a construction task. This study will provide insight into construction task modelling, multi-granularity task primitive learning, and visual interpretation of task primitive learning. The outcomes will enhance the mutual understanding of workers and robots and help improve the smoothness of a HRC team. In the following, we first review the related work in task primitive modelling, learning, and visual explanation in Section 2. Then, we design a multi-granularity task representation model and present the proposed method for learning multi-granularity task primitives in Section 3. We test our method and present the findings in Section 4. Finally, we conclude our study in Section 5.

2. Literature Review

This section reviews construction task modelling, vision-based construction task primitive learning, and visual explanation methods. The first and second aspects form the structure of a construction task and the technical premise for learning task primitives, respectively. The latter helps us understand what a robot learns from the visual world.

2.1 Construction Task Modelling

Task modelling involves efforts in the fields of construction and robotics. We first review the task representation model in the construction domain and then explore other task modelling methods in the robotics realm.

A widely used representation model in construction is the Work Breakdown Structure (WBS), which breaks down a project using a hierarchical structure. The lowest level, including one or a group of tasks called work packages, is the smallest unit of work. Though WBS has advantages in project-level management, its granularity is too coarse for a HRC task. Besides, the element in WBS focuses on planned outcomes rather than actions and thus contributes less to improving the robot's understanding of worker behaviour. Unlike WBS, Wu *et al.* (2022) developed a more fine-grained task representation model. They decomposed a construction task into four levels (i.e., tasks, subtasks, activities, and actions). This model is suitable for simple and sequential tasks such as bricklaying. However, this model is unsuitable for complex tasks, as these tasks cannot be described using four granularity levels precisely and can be executed in different orders.

Hierarchical Task Model (HTM) (Hayes and Scassellati, 2016) and knowledge graph (KG)-based model (Zheng *et al.*, 2022) in the robotics field provide viable solutions for these problems. HTM is a hierarchical, tree-like structure consisting of subtasks with different levels of abstraction. Each node in HTM is a subtask, and itself is a set of subtasks following sequential or concurrent orders. The leaf node of HTM is the atomic subtask, i.e., human actions. These features enable HTM to model different execution orders for complex construction tasks. The KG-based model further decomposes human actions into a graph-like representation, which encodes information about the attributes of interacting objects, the environment, and the agent. Though HTM and KG-based models can model various execution orders of a complex task using multi-granularity task primitives, these models are unsuitable for modelling construction tasks involving implicit cyclic processes. For example, in scaffolding construction, a worker may need to adjust the coupler and measure the levelness of the tube multiple times in order to reach an ideal position. However, this cyclic action pattern is uncertain and dependent on environmental conditions. This nature of construction tasks leads to implicit and dynamic task flows that render these static models ineffective. It also implies that action-level task primitives cannot accurately represent a construction task, which inspires us to find a more appropriate granularity of task primitives.

2.2 Vision-based Construction Task Primitive Learning

Existing vision-based methods for learning construction task primitives can be divided into rule-based and data-based methods.

Rule-based methods generally require that knowledge of task primitives be captured and defined in advance so that task primitives can be recognised via predefined rules. Early studies (Gong and Caldas, 2010) identified task primitives through the motion of large equipment, as the motion pattern is limited and can be easily predefined. After that, studies learn task primitives using spatial and temporal relations between workers and equipment. For example, Luo *et al.* (2018) recognised worker activities through predefined spatial patterns, such as equipment, materials, workers and equipment, and workers and materials. In addition to spatial relationships, Martinez *et al.* (2021) also considered temporal relationships. They used virtual finite state machines to model the spatial pattern between workers and equipment as well as the temporal pattern between task primitives. Though rule-based methods are well interpretable, they are challenged by the diverse construction activities and the dynamic environment when defining rules.

Data-based methods learn task primitives directly from visual data and can be roughly divided into handcrafted feature-based and ConvNet-based methods. The early study identified task primitives using handcrafted features such as the histogram of gradient and the histogram of optical flow (Gong, Caldas and Gordon, 2011). Though handcrafted feature-based methods require less training data (Lin *et al.*, 2020), these methods are limited to low-level features, such as edges and curves, which are not sufficient to characterise and distinguish complex worker action patterns (Zhang, Wang and Gao, 2021). With the evolution of deep learning, ConvNet-based methods have enabled unprecedented breakthroughs in task-primitive learning. For example, Luo *et al.* (2019) identified twelve low-level task primitives with a 3D ConvNet. Luo *et al.* (2020) recognised high-level task primitives using a hierarchical method, combining a 3D ConvNet and a conditional random field model. These studies have proven that ConvNet-based methods have great potential for task-primitive learning. However, most existing studies focus on a specific granularity level that is easily labelled. These studies are unable to gain a general understanding of the task in order to determine the best granularity for the task. Besides, these studies could not explain their predictions due to the black-box nature of neural networks. Thus, we expect a ConvNet-based approach that can learn multi-granularity task primitives as well as generate an explanation about the prediction.

2.3 Visual Explanation of ConvNets

ConvNet-based task primitive modelling methods sacrifice interpretability for accuracy. To increase the model's transparency, the visual explanation is the most straightforward approach (Wang *et al.*, 2019). It gains users' trust by highlighting the important regions influencing the predictions. Existing studies on the visual explanation of ConvNets include three methods: gradient visualisation, perturbation, and class activation map (CAM). Gradient-based methods generate a saliency map by backpropagating the gradient of a target class to the input layer (Simonyan and Zisserman, 2014). However, the quality of these maps is low (Omeiza *et al.*, 2019). Perturbation-based methods determine important input regions by perturbing the original input according to the observed changes in the model's prediction (Ribeiro, Singh and Guestrin, 2016). However, these methods are time-consuming to find the minimum region (Wang *et al.*, 2019). CAM-based methods generate saliency maps through a linearly weighted combination of activation maps (Wang *et al.*, 2019). These methods can provide high-quality and efficient visual explanations for a single input. Among CAM-based methods, gradient-weighted class activation mapping (Grad-CAM) (Selvaraju *et al.*, 2016) is the most widely adopted approach (Bao *et al.*, 2022), as it can produce saliency maps without altering ConvNet structure. Hence, Grad-CAM is adopted in this study to generate visual explanations.

2.4 Research Gaps

The literature review so far has identified three limitations on task primitive modelling and learning: (1) existing task representation models are unable to accurately define construction task structures involving implicit cyclic action patterns, affecting robots' task primitive learning and understanding; (2) there is a lack of multi-granularity task primitive learning methods that provide holistic information to determine which granularity is best suited for the task at hand. It is thus prone to error when manually defining the level of granularity, as it depends largely on individual knowledge and experience; and (3) there is a lack of transparency in ConvNet-based task primitive learning methods, and whether the model learned useful visual cues or dataset biases to solve the task remains unknown, hindering the establishment of a trustworthy HRC team.

3. Methodology

To bridge these gaps, this study proposes a ConvNet-based method for learning multi-granularity task primitives. Specifically, regarding task primitive modelling and learning, we first formulate the representation of multi-granularity task primitives and then design a Multi-Task Learning (MTL)-based ConvNet to classify task primitives at different granularities. Regarding the interpretation of model predictions, we visualise learned visual cues using Grad-CAM and expect the model to learn useful visual representations from construction videos.

3.1 Representation of Multi-granularity Task Primitives

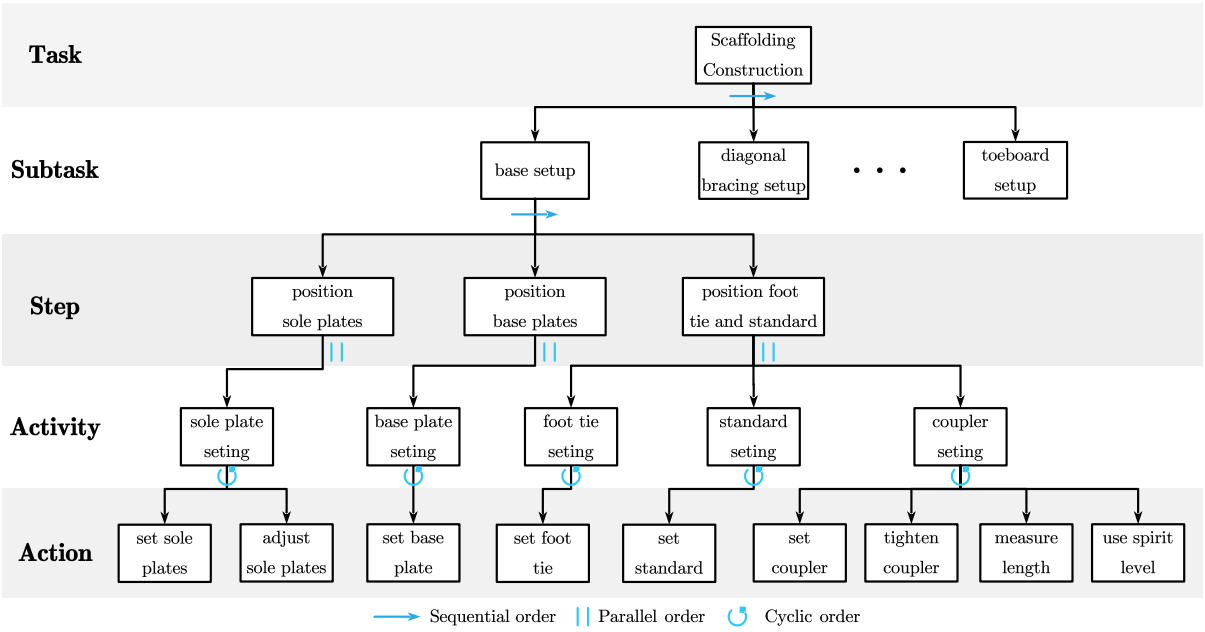


Fig. 1: A hierarchical decomposition of a scaffolding construction task. The task is represented as a tree of primitives of varying complexity and abstraction.

Previous studies (Cheng *et al.*, 2020; Wu *et al.*, 2022) concerned only simple tasks and developed a three-level task representation, i.e., tasks, subtasks, and actions (activities). However, this representation model is not suitable for modelling complex construction tasks due to the limited level of abstraction and the implicit, cyclic construction task flow. Hence, we propose two extra intermediate task primitives between subtasks and actions, i.e., steps and activities. This decomposition adheres to the WBS model's 100 percent rule, i.e., to model the outcome but not the process. Thus, high-level task primitives can be broken down into low-level task primitives with greater precision. Specifically, the step is the further decomposition of subtasks, and the activity is the abstraction of actions, which attempts to model the outcome of dynamic worker actions. Their definitions are shown as follows:

- **Action**: An action is a single worker's movement (e.g., walking) or complex interactions concerning the worker and their environment (e.g., carrying tubes and assisting other workers). Action is the finest-grained task primitive.
- **Activity**: Activity is the abstraction of actions, which consist of one or more actions. It represents the goal state or planned result of a sequence of actions, which can be executed cyclically.

- **Step**: A step is an element of completing a subtask and the goal state of a sequence of activities, which can be implemented with multiple sequences of activities.
- **Subtask**: A subtask is the basic unit of a task. It is achieved through a defined sequence of steps.
- **Task**: A task represents the work to be conducted by a HRC team. It specifies the initial state and the goal state. It can be decomposed into a set of fixed-order subtasks.

The hierarchical relationship of the above task primitives is shown in Fig. 1, which represents a real-world scaffolding construction task.

3.2 ConvNet Architecture

The proposed ConvNet architecture consists of two parts: the backbone and the classification head, as shown in Fig. 2. The backbone network is used to extract visual features from videos, and theoretically, it can be any ConvNet-based or vision transformer-based action recognition model. In this study, we chose the I3D model, which balances accuracy and efficiency well (Carreira and Zisserman, 2017). After the feature extraction performed by I3D, the extracted visual feature will be passed to the classification head for further processing. In the classification head, the multi-task learning (MTL) mechanism is adopted for multi-granularity task primitive recognition. MTL is a learning paradigm aiming to exploit useful information and learn a shared representation in related tasks to improve the generalisation performance of all tasks (Zhang and Yang, 2022). The intuition behind MTL is that related tasks can provide additional knowledge and serve as a regulation in joint training (Ruder, 2017). Following this spirit, we design a head with four classification tasks, i.e., action, activity, step, and subtask recognition, such that the efficiency of task primitive learning and the model's performance can be improved by using a single network and MTL. Specifically, the head is a collection of an average pooling layer, a dropout layer, and a fully connected (FC) layer. The number of neurons in the FC layer is the total of the four types of primitives. To train the model, four cross-entropy loss functions are introduced and function in different areas of the output of the FC layer. As we treat four tasks as equally important, the total loss is the sum of the losses from the four tasks with the same weighting.

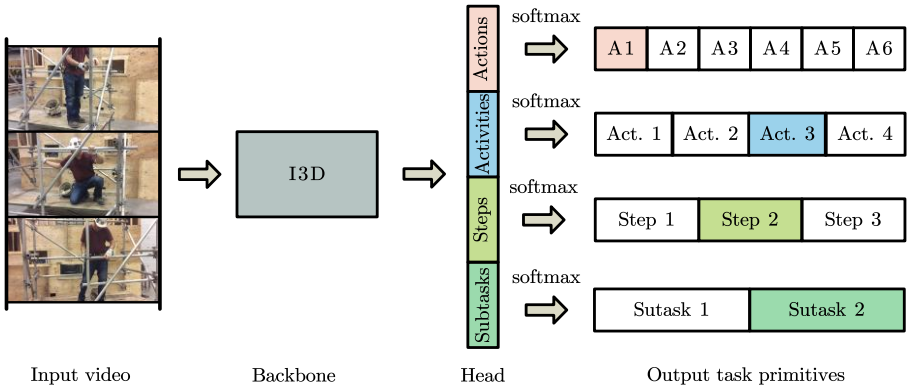


Fig. 2: The proposed ConvNet-MTL architecture. (change img upside down)

3.3 Grad-CAM

Grad-CAM is a visualisation technique that produces saliency maps to indicate important areas of input. It can efficiently generate high-quality heatmap for each input without changing the original network structure. This technique visualises learned visual features from the last

convolutional layer, as it assumes that features in this layer achieve a good balance between high-level semantics and spatial information. When given an image and a targeted class, Grad-CAM forward propagates the image via the CNN backbone and then through the task-specific head to obtain a raw score (e.g., the vector before the softmax layer). The gradient is set to zero for all classes except for the target class, which is set to one. This signal is then backpropagated until the last convolutional layer. A weighted vector is computed through a global average pooling for each channel of this backpropagated feature map. Finally, a localisation heatmap is generated via a weighted linear combination of a forward feature map and ReLU activation.

4. Experiments

To ground our study, a challenging scaffolding construction task was selected as an empirical study. This task requires multi-worker collaboration and has a high degree of uncertainty and complexity regarding worker actions and execution sequences. This section describes the experimental setup and findings.

4.1 Experimental Setup

Given the fact that RGB cameras are the most commonly used visual sensors in robots, seven videos concerning different subtasks of scaffolding construction have been collected. These videos were taken in an indoor environment with varying viewpoints. The total duration of these videos is approximately 40 minutes. According to the task representation in Section 3.1, we divided these videos into 9 subtasks, 13 steps, 16 activities, and 26 actions. We extract video clips from every 3 seconds of the video and create a dataset with 738 clips after removing the irrelevant clips. Each clip was manually annotated using these four granularity-level task primitives simultaneously. The training-to-test ratio is 4:1. To ensure a balance of actions across categories, we oversample actions with small sample sizes and undersample actions with large sample sizes. The statistics of the task primitives at the four granularities are shown in Table 1. The proposed model was implemented using PyTorch 1.8.1 and trained on an RTX 3070 GPU. This experiment adopts the warm-up training strategy with an initial learning rate of 0, gradually increasing to 0.0125 after 34 epochs. The total number of training epochs is 120, and the learning rate is reduced according to the cosine annealing policy. The optimiser is SGD, with a momentum of 0.9 and a weight decay of 0.00001.

Table 1: Statistics of task primitives in the scaffolding construction task. The number in brackets represents the number of video clips.

No.	Action		Activity		Step	Subtask
1.	set plan bracing (13)	secure mushroom coupler (35)	mushroom coupler setting (61)	guardrail setting (29)	install transom ledger (40)	putlog setup (39)
2.	cut wire (31)	set diagonal bracing (16)	sole plate setting (51)	tread board fixing (92)	install putlog (39)	ledger transom setup (40)
3.	set base plate (14)	use spirit level (29)	base plate setting (14)	tread board setting (19)	install toe board (82)	toe board setup (82)
4.	tighten coupler (56)	set ladder (26)	foot tie setting (18)		install plan bracing (14)	ladder setup (94)
5.	set foot tie (18)	climb up platform (34)	diagonal bracing setting (16)		install foot tie and standard (103)	working platform setup (222)

6.	use ladder (32)	adjust sole plate (34)	transom ledger setting (22)	prepare wire (31)	plan bracing setup (14)
7.	set putlog (30)	set transom ledger (22)	toe board setting (82)	install ladder (63)	guardrail setup (43)
8.	set coupler (13)	set mushroom coupler (28)	wire setting (31)	install tread board (19)	base setup (168)
9.	set sole plate (17)	secure tread board by wire (58)	ladder setting (95)	install diagonal bracing (35)	diagonal bracing setup (35)
10.	set toe board (33)	tighten toe board (17)	plan bracing setting (13)	install sole plates (51)	
11.	secure ladder (37)	set tread board (19)	standard setting (24)	secure tread board (153)	
12.	measure length (42)	set guardrail (29)	coupler setting (140)	install base plates (14)	
13.	use wire (32)	set standard (24)	putlog setting (30)	install guardrail (93)	

4.2 Multi-granularity Primitive Learning Results

The training and test results are shown in Fig. 3. The top-1 accuracy of four task primitives increases dramatically in the first 40 epochs. Though the increased magnitude of the subtask and step has since levelled off, the action and activity still increase, and the loss drops constantly. These results indicate that the model is not overfitting the dataset. The highest top-1 accuracy is achieved by the steps (1.00), then subtasks (0.99), activities (0.98), and actions (0.88). For understanding worker activities and the construction task, activity is optimal as it achieves the best compromise between accuracy and task knowledge. Compared to the step level and subtask level, the intermediate activity level has comparable recognition accuracy and conceals fewer subclasses, allowing it to reveal more specific task information. Although activity level cannot provide as detailed worker interactions as action level, it can produce more realistic sensing results due to the proper abstraction.

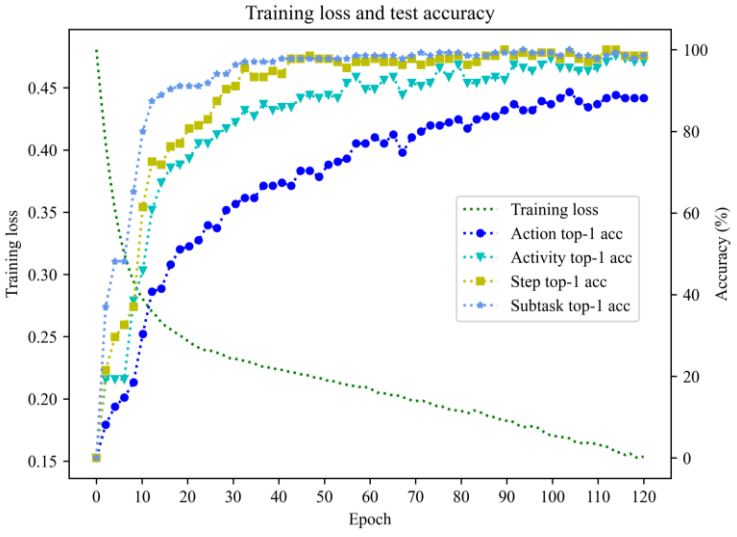
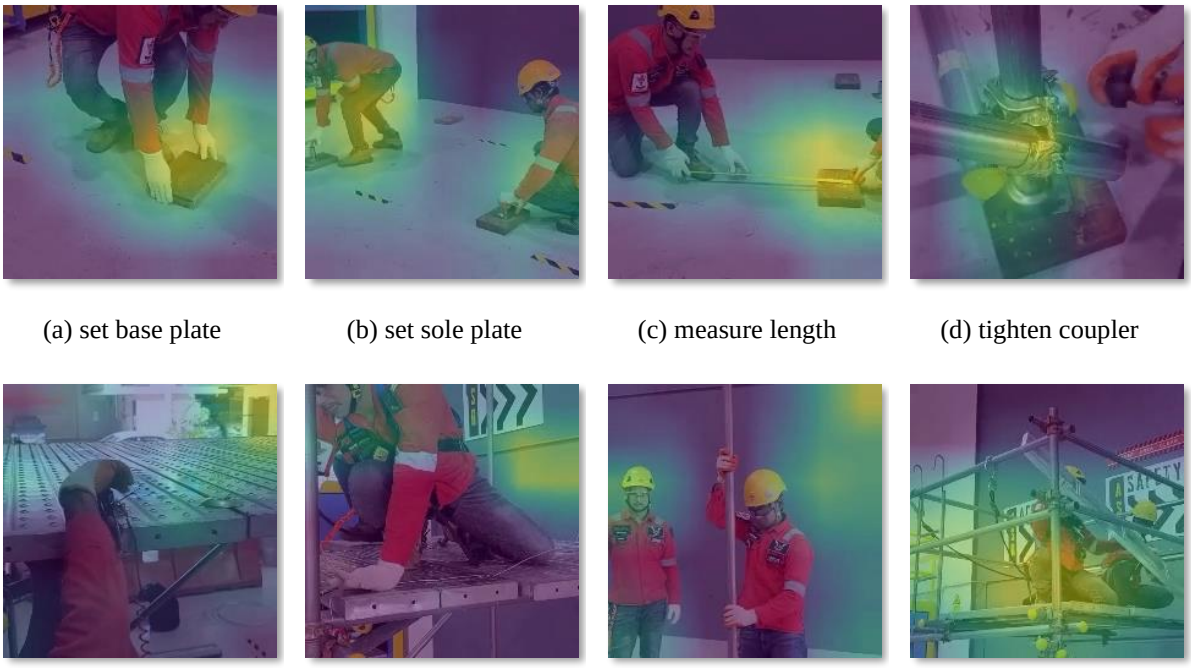


Fig. 3: Training loss and test accuracy at different epochs.

4.3 Visual Explanation Results

Using Grad-CAM, visual explanations with eight scenarios as examples are generated to determine whether the model learns useful visual cues or dataset biases. The results are shown in Fig. 4, where the region highlighted in yellow is supposed to be the focus of the model. The first row in Fig. 4. shows that the model learns useful cues when the background is simple and the workers' actions are obvious. However, when the background becomes complex or the worker's movement becomes less obvious, the model's focus shifts to the irrelevant background, as shown in the second row of Fig. 4. These results indicate that the model can learn useful cues in simple scenarios while also tending to learn dataset biases to distinguish task primitives in complex scenarios. To mitigate this problem, larger datasets and more advanced action recognition models should be explored, as they can improve models' generalizability. Besides, multi-modal data fusion methods and data pre-processing methods should also be investigated to acquire more cues and remove noise and interference from the input. Future applications are anticipated to transmit this visual explanation to the AR/VR device of a worker partner. Thus, the worker can assess the robot's decisions when the robot intends to provide assistance.



(a) set base plate (b) set sole plate (c) measure length (d) tighten coupler

(e) set mushroom coupler (f) set tread board (g) set standard (h) set toe board

Fig. 4: The visual explanation of learned visual cues.

5. Conclusion

Ensuring seamless HRC requires robots to learn at the proper level for a task and workers to make sense of the robots' perception process. Nevertheless, the implicit, dynamic construction task flow and massive interference in the captured visual data pose a challenge to achieving these goals. To address these challenges, we seek to enhance the mutual understanding between humans and robots by answering 1) which granularities of task primitives are best for the task at hand, and 2) whether the robot learns useful visual cues or dataset biases to solve the task.

The hierarchical representation of task primitives is first formulated to model the implicit, cyclic task flow, then an MTL-based task learning model is developed to learn four task primitives, and finally, the learned visual cues of the model is visualised using Grad-CAM. Using

scaffolding construction as a case study, this study found that intermediate-level activity is the optimal granularity because it achieves the best balance between recognition accuracy and the level of detail of task knowledge. This study also found that the model learned both useful and useless visual cues to separate task primitives in different scenarios. The research contributes to the body of knowledge by providing a transparent approach for modelling dynamic construction tasks and recognizing multi-granular worker behavior from construction videos. The outcomes will improve the mutual understanding between workers and robots and help facilitate a seamless HRC.

6. References

Anjomshoae, S. *et al.* (2019) "Explainable Agents and Robots: Results from a Systematic Literature Review," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems (AAMAS' 19), pp. 1078–1088.

Bao, H. *et al.* (2022) "Multi-granularity visual explanations for CNN," *Knowledge-Based Systems*, 253, p. 109474. doi: https://doi.org/10.1016/j.knosys.2022.109474.

Bock, T. and Linner, T. (eds.) (2016) "Single-Task Construction Robots by Category," in *Construction Robots: Elementary Technologies and Single-Task Construction Robots*. Cambridge: Cambridge University Press, pp. 14–290. doi: DOI: 10.1017/CBO9781139872041.002.

Brosque, C. *et al.* (2020) "Human-Robot Collaboration in Construction: Opportunities and Challenges," in *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–8. doi: 10.1109/HORA49412.2020.9152888.

Carreira, J. and Zisserman, A. (2017) "Quo Vadis, Action Recognition? {A} New Model and the Kinetics Dataset," *CoRR*, abs/1705.0.

Cheng, Y. *et al.* (2020) "Towards Efficient Human-Robot Collaboration With Robust Plan Recognition and Trajectory Prediction," *IEEE Robotics and Automation Letters*, 5(2), pp. 2602–2609. doi: 10.1109/LRA.2020.2972874.

Gong, J. and Caldas, C. H. (2010) "Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations," *Journal of Computing in Civil Engineering*, 24(3), pp. 252–263. doi: 10.1061/(ASCE)CP.1943-5487.0000027.

Gong, J., Caldas, C. H. and Gordon, C. (2011) "Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models," *Advanced Engineering Informatics*, 25(4), pp. 771–782. doi: https://doi.org/10.1016/j.aei.2011.06.002.

Grigore, E. C. *et al.* (2018) "Preference-Based Assistance Prediction for Human-Robot Collaboration Tasks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4441–4448. doi: 10.1109/IROS.2018.8593716.

Hayes, B. and Scassellati, B. (2016) "Autonomously constructing hierarchical task networks for planning and human-robot collaboration," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5469–5476. doi: 10.1109/ICRA.2016.7487760.

Lin, W. *et al.* (2020) "Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment," *Scientific Reports*, 10(1), p. 20336. doi: 10.1038/s41598-020-77264-y.

Liu, Y. and Jebelli, H. (2022) "Intention Estimation in Physical Human-Robot Interaction in Construction," *Construction Research Congress 2022: Computer Applications, Automation, and Data Analytics, CRC 2022*. American Society of Civil Engineers (ASCE), pp. 621–630. doi: 10.1061/9780784483961.065.

Luo, X. *et al.* (2018) "Recognising Diverse Construction Activities in Site Images via Relevance Networks of Construction-Related Objects Detected by Convolutional Neural Networks," *Journal of Computing in Civil Engineering*, 32(3), p. 4018012. doi: 10.1061/(ASCE)CP.1943-5487.0000756.

Luo, X. *et al.* (2019) "Vision-based detection and visualisation of dynamic workspaces," *Automation in Construction*, 104, pp. 1–13. doi: https://doi.org/10.1016/j.autcon.2019.04.001.

Luo, X. *et al.* (2020) "Combining deep features and activity context to improve recognition of activities of workers in groups," *Computer-Aided Civil and Infrastructure Engineering*, 35(9), pp. 965–978. doi: 10.1111/mice.12538.

Martinez, P. *et al.* (2021) "A vision-based approach for automatic progress tracking of floor paneling in offsite construction facilities," *Automation in Construction*, 125, p. 103620. doi: https://doi.org/10.1016/j.autcon.2021.103620.

Omeiza, D. *et al.* (2019) "Smooth Grad-CAM++: An Enhanced Inference Level Visualisation Technique for Deep Convolutional Neural Network Models," *CoRR*, abs/1908.0.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) " 'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery (KDD' 16), pp. 1135–1144. doi: 10.1145/2939672.2939778.

Ruder, S. (2017) "An Overview of Multi-Task Learning in Deep Neural Networks," *CoRR*, abs/1706.0.

Selvaraju, R. R. *et al.* (2016) "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization," *CoRR*, abs/1610.0.

Simonyan, K. and Zisserman, A. (2014) "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, abs/1409.1.

Wang, H. *et al.* (2019) "Score-CAM: Improved Visual Explanations Via Score-Weighted Class Activation Mapping," *CoRR*, abs/1910.0.

Wu, H. *et al.* (2022) "A survey on teaching workplace skills to construction robots," *Expert Systems with Applications*, 205, p. 117658. doi: https://doi.org/10.1016/j.eswa.2022.117658.

Zhang, J., Wang, P. and Gao, R. X. (2021) "Hybrid machine learning for human action recognition and prediction in assembly," *Robotics and Computer-Integrated Manufacturing*, 72, p. 102184. doi: 10.1016/j.rcim.2021.102184.

Zhang, Y. and Yang, Q. (2022) "A Survey on Multi-Task Learning," *IEEE Transactions on Knowledge and Data Engineering*, 34(12), pp. 5586–5609. doi: 10.1109/TKDE.2021.3070203.

Zhang, Z. *et al.* (2022) "Prediction-Based Human-Robot Collaboration in Assembly Tasks Using a Learning from Demonstration Model," *Sensors* . doi: 10.3390/s22114279.

Zheng, P. *et al.* (2022) "A visual reasoning-based approach for mutual-cognitive human-robot collaboration," *CIRP Annals*, 71(1), pp. 377–380. doi: 10.1016/j.cirp.2022.04.016.