Hierarchical Fine-Grained Action Recognition for Manufacturing Applications

Phan Trong Huy

OKI is conducting research and development into image sensing to solve a variety of mission-critical issues, including those related to people/objects at manufacturing sites and traffic incidents on roads. At manufacturing sites where manual work remains essential, OKI is developing worker action recognition technology using image sensing to achieve higher levels of quality and process management. For example, the technology will be extremely useful in confirming correct work procedures on a production line and analyzing work times[1), 2)]. The technology extracts information useful for a specific application, such as the target person's skeletal structure, from video images, and then processes the information using a time-series deep learning model to classify a sequence of detailed actions.

This article introduces the hierarchical classification technology as one method of improving the performance of action recognition. To validate the technology's effectiveness, a test was performed using a dataset[3)] depicting an assembly of desktop PCs in a manufacturing setting. The result of the test is presented in this article as well.

Hierarchical Classification Technology

- **Current Technical Issues**

  Research into the detailed classification of human actions in video images is attracting attention, and applications are progressing in various fields, including healthcare and manufacturing.

  Human action recognition using image sensing can be categorized into "coarse-grained action recognition" and "fine-grained action recognition." In coarse-grained action recognition, the entire human body is captured on video, and general actions such as "running," "jumping," and "sitting" are recognized. This type of action classification is achieved by using "visible" information from the person or background, and overall "movement" information. On the other hand, fine-grained action recognition is a relatively difficult task that focuses primarily on the movement of a person's arms and parts of the body to classify a sequence of detailed actions related to a task, such as "picking up an object" or "putting down an object"[2)].

  Classifying "jumping," "running," "playing sports," and other coarse-grained actions found in everyday life, where the background and human movements vary significantly, can be treated as a task that classifies a number of actions. Whereas, sequence of actions such as "standby," "attach legs to table," and "rotate legs" during table assembly at a manufacturing site have different granularity for indicating the classification level and the boundaries can be ambiguous. This makes flat learning difficult and results in mutual misclassification. As such, when there

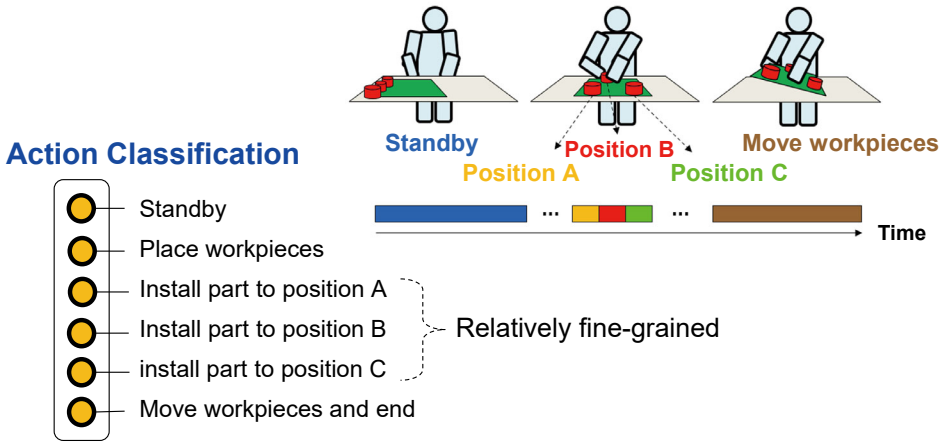


Figure 1. Flat Classification of Sequence of Actions During Assembly

are numerous target actions that occur in succession, the boundaries become ambiguous, making it difficult to learn and classify all actions in the same flat manner as in the past.

- **Hierarchical Classification**

Sequence of actions can have varying levels of granularity depending on the annotation. In the example shown in **Figure 1**, the three fine-grained actions of "install position-specific part" are difficult to learn since they must be distinguished not only from each other but also from coarse-grained actions such as "standby." Therefore, to improve classification performance, OKI developed a technology that appropriately groups the target actions based on their granularity. A single task that was previously difficult to classify is hierarchically divided into multiple tasks that are easy to classify.

Actions are classified coarsely at the upper level and finely at the lower level. Actions that only represent a person's state (e.g., standby) are classified coarsely, while actions that represent similar actions on objects (e.g., pick up part X, pick up part Y) are classified more finely. How specific actions are performed is divided into detailed groups at a level that is further lower (e.g., tighten screw in position A, tighten screw in position B). With this method, if the annotated actions are fine-grained, it will be necessary to create new representative action classes (e.g., work in progress, tightening screw) that belong to a higher, coarser level.

In the assembly example, the original flat target actions are divided into three levels of different granularity, as shown in **Figure 2**, and classification learning and inference are performed. The three actions with the finest granularity and similar movements, "install part to position A/B/C," are grouped together, and the representative of these, "install part," is placed at a coarser level. By limiting the number of actions to be classified in each group to two or three, a more balanced learning is possible. Furthermore, hierarchically performing action inference prevents misclassification with actions in other levels or groups. For example, with flat classification, "install part to position C" may be mistaken for the next action, "move workpieces," while with hierarchical classification, misclassification only occurs with position A or B. Therefore, when detailed determination such as the position of a part installation is difficult, the method has the advantage of outputting the appropriate inference result of "part installation." Moreover, it can lead to the discovery of non-standard work and other actions (**Figure 3**).

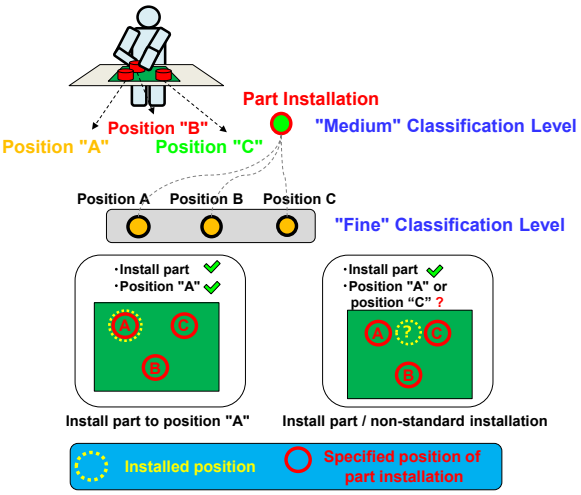


Figure 3. Example of Utilizing Hierarchical Classification Results

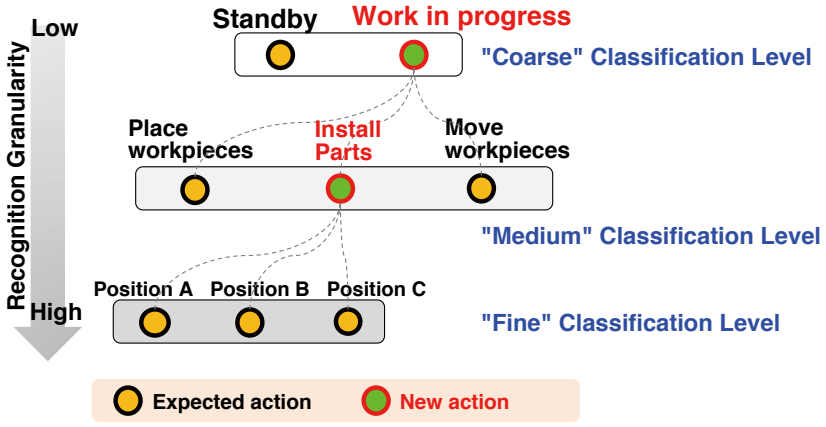


Figure 2. Hierarchical Classification of Sequence of Actions During Assembly

Validation Details

- **Validation Dataset**[3)]

The desktop PC assembly task used in the validation test involved the subject to install multiple components laid out on the tabletop, including a fan, RAM, and SSD, into the PC body (**Figure 4**). Depending on the component, a screwdriver and screws were also used during installation. The actions to be classified were annotated with detailed information, such as “Pick_up_RAM/Install_RAM” (action and target object) and “Tighten_screw_1/Tighten_screw_2” (action, target object, position).

In the hierarchical classification described in this article, action classes were organized into a total of six groups in three hierarchies, as shown in **Figure 5**, and learning and inference were performed.

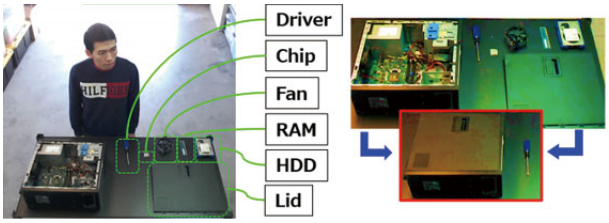


Figure 4. Desktop PC Assembly Task

Validation and Analysis Results

The validation test was performed using a time-series deep learning model and skeletal information. As shown in **Table 1**, hierarchical classification demonstrated overall improved performance compared to flat classification. This is likely due to the small number of similar level actions per each level, resulting in a more proper learning. With all 20 flat classes, misclassification is likely to occur when similar actions, such as “Pick_up” and “Place/Install” actions are performed on target parts that are located close to each other. The same tendency for misclassification is also observed when a more detailed position-specific classification is required, such as the “Tighten_screw” action. On the other hand, hierarchical classification avoids misclassification by dividing tasks, with higher levels responsible for general classification and lower levels responsible for detailed classification, as shown in **Figure 6**. For example, in the “Pick_up_chip” action, the hierarchical method classifies something is picked up during the “In_progress” action and that the target part is a chip during the “Pick_up” action. In this way, the hierarchical method classifies the detailed actions involved in assembling a desktop PC with greater accuracy than the conventional flat method, making it possible for the hierarchical method to be used in applications that detect procedural errors and measure work time for each step.

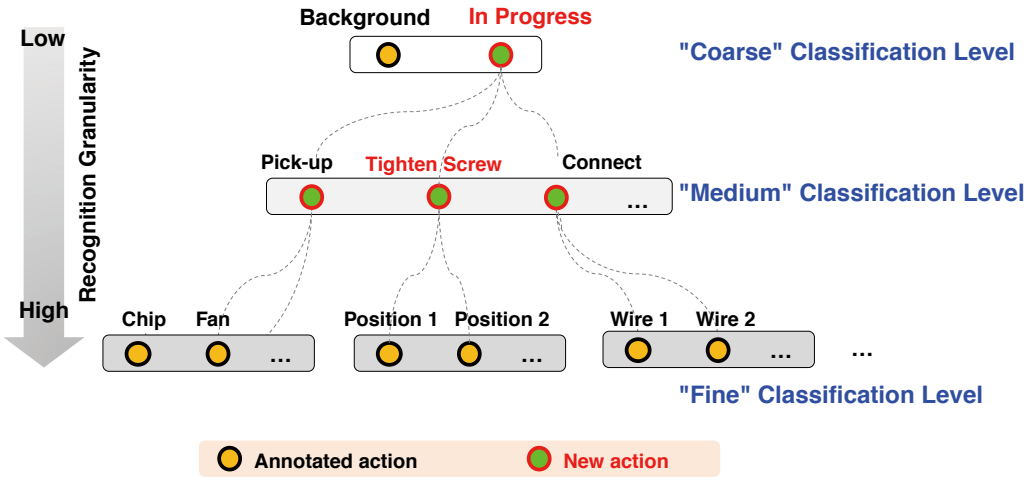
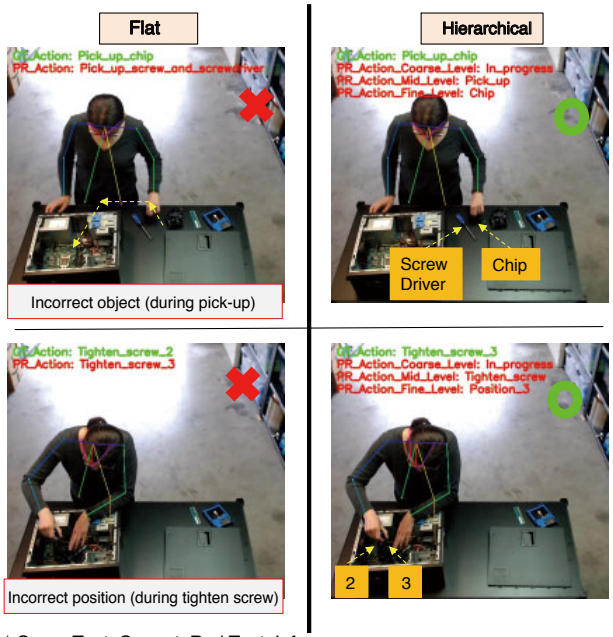


Figure 5. Hierarchical Classification of Desktop PC Assembly Data

Table 1. Evaluation Results of Frame Accuracy (%)

Level/Group	Action Class	Hierarchical	Flat
Coarse Classification (2 classes)	In Progress	99.5	–
	Background	92.6	92.7
Medium Classification (6 classes)	Pick_up	89.4	–
	Place/Install	91.2	–
	Tighten_screw	96.8	–
	Put_down (Screwdriver)	82.9	76.9
	Connect	92.6	–
	Close (Lid)	98.6	96.8
Fine Classification Group 1 (Pick up actions /6 classes)	Pick_up_chip	96.9	81.8
	Pick_up_screw_and_driver	97.3	79.3
	Pick_up_fan	90.3	74.7
	Pick_up_RAM	90.4	74.8
	Pick_up_HDD	88.7	77.8
	Pick_up_lid	95.1	82.3
Fine Classification Group 2 (Place/Install actions /4 classes)	Place_chip_on_motherboard	95.9	88.3
	Place_fan_on_motherboard	94	68.8
	Install_RAM	95.4	90.7
	Install_HDD	99.3	86.8
Fine Classification Group 3 (Tighten screw actions /4 classes)	Tighten_screw_1	92	86.1
	Tighten_screw_2	94	88
	Tighten_screw_3	93.1	93
	Tighten_screw_4	97.8	93.2
Fine Classification Group 4 (Connect actions /3 classes)	Connect_wire_to_motherboard	98.2	84.3
	Connect_wire_1_to_HDD	91.7	83.1
	Connect_wire_2_to_HDD	89.9	76



* Green Text: Correct, Red Text: Inference
* Hierarchical classification displayed in three levels (Coarse_Level, Mid_Level, Fine_Level)

Figure 6.
(Top) Classification Results for Pick-Up Part Action
(Bottom) Classification Results for Tighten Screw by Position Action

Conclusion and Future Outlook

This article introduced the hierarchical classification technology as one method of improving the performance of worker action recognition. The technology's effectiveness was validated through testing using a dataset depicting an assembly of desktop PCs in a manufacturing setting.

OKI is currently continuing the effort to refine the fine-grained action recognition technology so that it can be applied to a variety of tasks and applications at manufacturing sites where manual work is essential. Going forward, OKI is planning research and development on achieving advanced modeling, such as "task analysis" to uncover unnecessary work. ◆◆

References

1) Ryoya Kawatsura, Masahito Asano, Kouji Araragi, Kazuma Yamamoto, Tsukasa Kobayashi: Behavior Judgment System for Justifying Correctness of Work Contents and Processes, OKI Technical Review, Issue 238, Vol. 88 No. 2, pp.46-49, November 2021 (in Japanese)
2) Phan Trong Huy, Takayuki Ueno, Kazuma Yamamoto: Fine-Grained Worker's Action Recognition for Manufacturing Sites, OKI Technical Review, Issue 239, Vol.89 No.1, pp.16-19, May 2022 (in Japanese)
3) The Desktop Assembly Dataset https://retrocausal.ai/

Authors

Phan Trong Huy, AI R&D Department, Research & Development Center, Technology Division

TIPS【Glossary】

Annotation
Process of adding tags and metadata to data such as text, audio, images, and video to make the information easier to understand. Particularly in the AI field, it is the process of creating training data (correct answer data) for machine learning models, and it is essential for improving AI performance.