S & M 4382

Task Time Estimation Based on Action Classification in Dyeing Processes Using Video Data

Kazuma Sakamoto,[1*] Fuya Shibata,[2] Iori Iwata,[2] Aki Mimura,[3] and Yoshihiro Ueda[1]

[1]Faculty of Production Systems Engineering and Sciences, Komatsu University,
Komatsu, Ishikawa 923-8511, Japan
[2]Graduate School of Sustainable Systems Science, Komatsu University, Komatsu, Ishikawa 923-8511, Japan
[3]KOMATSU MATERE Co., Ltd., Nyu-167, Hamamachi, Nomi, Ishikawa 929-0124, Japan

(Received December 1, 2025; accepted February 13, 2026)

Keywords: VGG16, ST-GCN, two-stage model, classification, task time estimation

In this research, we developed a two-stage model that integrates first-person video and third-person skeletal information to perform task classification and task time estimation in a textile dyeing process. In the first stage, Visual Geometry Group 16 (VGG16) recognizes task-related objects, whereas in the second stage, Spatial Temporal Graph Convolutional Network (ST-GCN) classifies detailed task actions. Through the incorporation of additional training data for the latter stage, we have attained enhanced classification accuracy across all categories, with particularly noteworthy advancements observed in tasks characterized by ambiguous boundaries. Furthermore, by automatically detecting task start and end points from the predicted action label sequences, we estimated task durations and confirmed reductions in estimation error through noncontact sensing using cameras as optical sensors combined with advanced action recognition technologies. Moreover, by calculating the mean and standard deviation of task durations for each task, we were able to evaluate process bottlenecks and the stability of each task. The proposed method demonstrates the feasibility of achieving task analysis and time estimation through noncontact, camera-based measurement without disrupting on-site operations. This approach offers a promising framework for facilitating process optimization and task standardization in manufacturing environments.

1. Introduction

In the manufacturing industry, the measurement of task time is imperative for evaluating productivity, establishing standard times, and identifying bottlenecks. In the context of designing human-centric production systems, it is imperative to consider factors such as worker workload and safety.(1) The accurate measurement and analysis of task time contribute to the optimization of production efficiency, the improvement of processes, and the playing of a critical role in achieving sustainable manufacturing through labor and energy reduction.(2) Conventionally, the measurement of task time has been conducted manually through the utilization of stopwatches and time-study sheets.(3) However, these methods impose significant burdens on observers,

[*]Corresponding author: e-mail: kazuma.sakamoto@komatsu-u.ac.jp

https://doi.org/10.18494/SAM6090

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

rendering them unsuitable for long-term continuous measurements or large-scale data collection. Additionally, the subjective assessment of the observer can affect the measurement outcomes. These limitations have prompted the development of automated methods that utilize camera images and sensor data to recognize work states and estimate task duration.

Beyond reducing the burden on observers, the automated analysis of dyeing processes aims to improve the work environment for the operators themselves. By identifying operational bottlenecks and eliminating redundant or high-strain movements, in this research, we proposed a two-stage framework for task time estimation to facilitate the standardization of work procedures. Such improvements are expected to simplify task execution and mitigate the physical workload on workers, leading to a more sustainable and worker-friendly production system.

On the other hand, certain potential drawbacks must be addressed for practical implementation. Continuous video recording can impose a psychological burden on operators and raises significant privacy considerations. Furthermore, a technical constraint of camera-based systems is the limited field of view. To ensure that tasks are not missed or obscured by large machinery, it is often necessary to install multiple cameras to cover the entire workspace. This increased number of devices may inadvertently intensify the workers' feeling of being under surveillance. In this study, we mitigated these issues by strictly adhering to ethical protocols, including obtaining informed consent and ensuring that the participants fully understood the research objectives before the experiments commenced.

The focus on dyeing processes in this research is motivated by their unique operational characteristics and critical role in textile manufacturing. First, task durations in dyeing exhibit significant variability—even for identical tasks—owing to factors such as worker skill level, fabric volume, and equipment configuration. Quantifying this variability is essential for process improvement and the design of human-centric production systems. Second, many dyeing facilities operate on a 24/7 basis and must frequently adapt to urgent orders or process contingencies. In such dynamic environments, reliable task time estimation is crucial for identifying bottlenecks that directly impact throughput and lead times. Furthermore, unlike the well-studied assembly processes in previous literature, dyeing environments are characterized by frequent occlusions caused by large machinery and materials, making video-based analysis more challenging. Demonstrating that task time estimation is feasible using only cameras under these real-world conditions is a key objective of this research.

In our previous work, we proposed a two-stage model for task classification in the textile dyeing process.(4) This model integrates a first-person video with third-person skeletal information. Specifically, VGG16 was employed to classify work objects in first-person videos, whereas MMpose and Spatial Temporal Graph Convolutional Network (ST-GCN) were utilized to estimate poses and analyze temporal dynamics in third-person videos, respectively, with the objective of identifying detailed work actions associated with each object.(5–7) The findings indicated that the proposed two-stage model demonstrated superior classification accuracy compared with using first-person or third-person information alone.

However, previous research(4) showed misclassification rates of approximately 10–20% for certain categories, particularly for similar actions, highlighting variability in classification performance. Additionally, task time estimation derived from the output of the two-stage model remained at a conceptual level, with no quantitative evaluation performed.

The objective of this research is to augment the existing literature by implementing task time estimation based on the output of the two-stage model and quantitatively evaluating its accuracy by comparing it with manually measured task times. Specifically, the first-person model estimates the task of the manipulated object, whereas the third-person model classifies detailed work actions at the frame level. The automated estimation of task duration relies on detecting the start and end times of the task.

Furthermore, a retraining experiment was conducted by augmenting the training dataset for the second-stage model (third-person model) in comparison with previous research. The impacts of data augmentation and increased sample size on action classification performance and task time estimation accuracy were evaluated. This approach enables high-precision, stable action recognition and task time estimation in the dyeing process by leveraging cameras as noncontact sensors, thereby contributing to the development of human-centric and sustainable manufacturing processes.

This approach demonstrates that the integration of visual data and skeletal dynamics functions as a robust sensing technology capable of capturing complex human motions in industrial settings without physical interference.

While in our previous work, we proposed the basic architecture of the two-stage model, its primary focus was on the feasibility of action classification. In the present work, we advance this research significantly by bridging the gap between classification and practical task time estimation. The main contributions of this paper are threefold.

First, we extend the two-stage framework to an automated task time estimation system and provide the first quantitative validation using on-site measurements. Unlike previous conceptual discussions, we implemented a rigorous evaluation pipeline that includes the automatic detection of task boundaries (start/end points), the calculation of relative estimation errors, and the statistical analysis of task durations (mean and variance). This allows for a granular understanding of the system's reliability in a real-world factory environment.

Second, we provide an experimental analysis of the relationship between training data volume and estimation accuracy. By conducting an ablation research study on the amount of training data for System 2 (ST-GCN), we clarify to what extent classification performance improvements contribute to the precision of time estimation. This provides a practical guideline for the amount of data required to deploy such systems in other manufacturing processes.

Third, we demonstrate the robustness of the system using multi-view video data (first-person and third-person) recorded simultaneously during the same work sessions. Previous studies utilized datasets recorded on different days, which did not account for the temporal and environmental consistency required for real-time synchronization. Validating the model with same-day recordings ensures its applicability to practical factory deployment where multi-view data is acquired in parallel.

2. Related Work

A substantial body of research has been dedicated to task time estimation and action recognition, with a particular focus on the manufacturing sector. Conventional methods have relied on manual measurements obtained through the use of stopwatches and work observations. However, these approaches impose considerable burdens on observers and are limited in terms of scalability and accuracy. Consequently, there has been a surge of interest in automated techniques that employ computer vision and sensor-based recognition to estimate task duration and identify work states.

Nakano and Shida developed a deep learning framework for the analysis of camera footage collected in an assembly factory.(8) Their research focused on the assembly processes, specifically the extraction of regions of interest around the hands and tools of the worker. Convolutional neural networks were employed to classify the ongoing task. By focusing on localized image patches that capture task-related object manipulations, their approach achieved a classification accuracy of 99.5% in a real-world factory setting, enabling high-reliability automated task time measurement based on frame-level task labels. However, the efficacy of this method is contingent on stable lighting conditions and the visibility of the worker's hands, which may not be met in environments characterized by frequent occlusions, such as textile dyeing processes.

Hitachi Ltd. proposed a method for estimating task duration based on worker dwell time within predefined zones, utilizing object detection and tracking to monitor worker positions. (9) Since this approach does not rely on precise action boundaries, it can be applied to tasks where the start and end points are ambiguous. However, its performance is heavily dependent on camera placement and the definition of work areas, making it difficult to distinguish between different actions performed within the same area.

A substantial body of research has sought to integrate action recognition with task-time estimation. Murai *et al.* developed a system that compares deep-learning-based action estimates with predefined standard operating procedures to detect deviations and time discrepancies in real time.(10) While this method is effective for structured assembly tasks, it assumes the availability of well-defined workflows and does not address environments prone to occlusions.

Matin *et al.* developed a hybrid deep-learning pipeline that integrates object detection with 2D pose estimation to recognize assembly actions and estimate task duration.(11) The proposed methodology involves explicitly tracking hand-object interactions, leveraging pose features to distinguish fine-grained actions. While this method is effective for tasks performed in well-lit environments, it requires high-quality multi-view inputs and controlled settings.

Li *et al.* proposed a real-time system for process progress estimation and phase detection in sequential processes.(12) The system utilizes a multimodal deep learning architecture to extract spatiotemporal features and a deep regression module to estimate process completeness from multiple sensor inputs. The estimated completeness is then used to predict the remaining time for continuous tasks in medical trauma resuscitation and Olympic swimming scenarios, highlighting the promise of learning-based time estimation in these domains. However, it does not address discrete manufacturing tasks with object-dependent action categories.

Moreover, recent research has placed increasing emphasis on the design of human-centric production systems. In their seminal work, Grosse *et al.* underscored the importance of integrating human factors into industrial work analysis.(13) This integration involved combining pose-estimation-based motion capture with well-established ergonomic assessment methodologies, such as the Rapid Upper Limb Assessment method. Their research indicates that AI-based posture estimation can be used not only for efficiency analysis but also for evaluating operator workload, fatigue, and musculoskeletal risks. These factors are pivotal in Industry 5.0 and human–cyber–physical production systems. This perspective aligns with the objective of the present research, which uses noncontact camera sensing for action recognition and task time estimation in real manufacturing environments. This approach contributes to human-centric and data-driven process optimization.

Wanyan *et al.* conducted a comprehensive survey of few-shot action recognition (FSAR).(14) The objective of this research was to address the high cost and impracticality of manually annotating large-scale video datasets for action recognition. The authors emphasize that, in comparison with few-shot learning in image or text domains, FSAR must contend with the inherent complexity of video data, including long-range temporal dependencies, rich semantic context, and significant intra-class variance arising from diverse appearance and motion patterns. The survey methodically categorizes existing FSAR methods into two overarching categories: generative approaches and meta-learning-based frameworks. Additionally, it provides a concise overview of the most commonly used benchmark datasets and offers a critical analysis of advanced topics and potential future research directions. This line of research underscores the importance of developing sample-efficient action recognition models in scenarios where annotation resources are scarce. This phenomenon is particularly pronounced in industrial environments, such as textile dyeing, where the collection and labeling of large-scale, task-specific video data can impose significant financial constraints. However, the survey's primary focus on generic video benchmarks excludes the direct consideration of task time estimation and the integration of first-person and third-person views in real factory processes, which are central themes of the present work.

Notwithstanding these advancements, a considerable number of existing studies employ fixed overhead cameras and operate under the assumption of stable viewpoints with minimal occlusion. Such assumptions are ill-suited for environments such as textile dyeing processes, where equipment, materials, and worker postures frequently obstruct visibility. Moreover, there is a paucity of studies that have employed both first-person and third-person viewpoints to integrate object-level context with worker motion patterns. To the best of our knowledge, no studies have concurrently addressed high-accuracy action recognition and task time estimation.

In contrast to previous studies, we introduce in this research a two-stage model that integrates first-person object-centric information with third-person skeletal motion features. The initial model identifies the work objects being manipulated from wearable-camera footage, whereas the subsequent model utilizes 2D skeleton sequences extracted from fixed-view factory cameras to classify detailed work actions. The proposed method combines both perspectives, enhancing the robustness of action recognition in occlusion-prone environments and enabling the frame-level detection of task start and end times. Furthermore, in this research, we expand the training dataset for the second-stage model (third-person model) and quantitatively evaluate the impact of dataset size on both action classification performance and task time estimation accuracy—an aspect that was not previously addressed in the literature.

3. Proposed Method

In this research, we investigated the dyeing process in a textile factory. Following established protocols, data were collected from both first-person and third-person perspectives. First-person videos were captured via a chest-mounted wearable camera, whereas third-person videos were recorded using a fixed camera positioned in front of the dyeing machine.

We propose a two-stage model to analyze these dual-perspective videos. This design was adopted because accurate task time estimation requires the stable classification of fine-grained categories, a task where single-view models often struggle as the number of categories increases. Specifically, while the first-person video clearly captures the objects being manipulated (e.g., dyeing machines, sewing machines, and carts), different actions performed on the same object often appear visually similar, leading to classification errors. Conversely, third-person videos effectively capture motion patterns through skeletal dynamics but lack sufficient object-related cues owing to occlusions and distance.

To address these limitations, we decompose the recognition process into two stages: the first-stage model analyzes first-person video for object classification and the second-stage model analyzes third-person video for action classification. For instance, if the first stage identifies the “dyeing machine”, the second stage then classifies specific actions such as “panel operation” and “opening/closing the lid”. Figures 1 and 2 illustrate the first-person and third-person perspectives used in this research, respectively.

3.1 Analysis of first-person perspective video data (first-stage model)

In this research, first-person perspective video footage is captured using a wearable camera (Driveman BC-100). The proposed method utilizes image recognition techniques to analyze first-person perspective videos of the dyeing process, thereby facilitating task classification. The system is composed of two primary components: the “task feature learning function” and the “task classification function”.

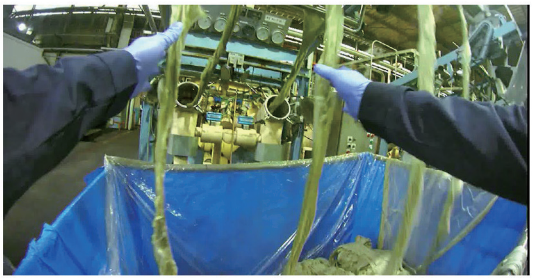

Fig. 1. (Color online) First-person perspective sample tasks.

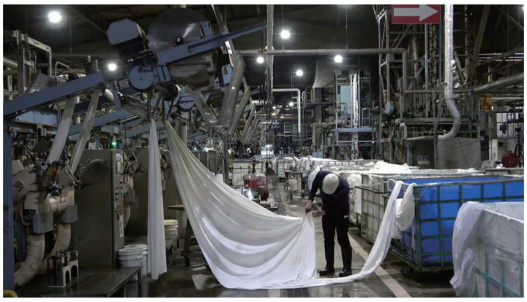

Fig. 2. (Color online) Third-person perspective sample tasks.

3.1.1 Task feature learning function (VGG16)

The task feature learning function is based on the VGG16 model pretrained on ImageNet, and a model specialized for task classification is constructed.(15) This function extracts visual patterns from the task data and provides the necessary information for the subsequent task classification function. The input data for this function consists of labeled training videos collected from the dyeing process. Feature learning is performed through the following procedure: First, the video is divided into individual frames, with each frame being converted into a format suitable for VGG16. Then, the pretrained VGG16 model is used to extract visual features from each frame, such as those corresponding to dyeing or sewing machines, and learn the objects and action patterns within the dyeing process. Finally, on the basis of the extracted features, a task classification model is built.

3.1.2 Task classification function (VGG16)

The task classification function uses the trained task classification model to classify evaluation data. In this function, the evaluation data is extracted from a dataset different from the training data and is used to assess the generalization performance of the model. Specifically, the evaluation data is first divided into individual frames, which are then transformed into an appropriate format, similar to the process during training. The preprocessed frames are subsequently input into the task classification model, where the features of each frame are analyzed to predict the task.

3.2 Analysis of third-person perspective video data (second-stage model)

The system consists of three main components: the pose estimation function, the task feature learning function, and the task classification function. In this method, pose estimation is performed on the video footage of the dyeing process captured by cameras installed within the factory. The skeletal information and temporal changes are then utilized for task classification.

3.2.1 Skeletal estimation function

In the pose estimation function, MMpose is used to extract skeletal information from the input video footage of the worker. The skeletal information consists of 17 keypoints, each represented as a 2D coordinate with respect to the top-left corner of the image. The list of keypoints used in this method is shown in Table 1. Keypoints corresponding to the face, ID1 to ID4, are excluded from the analysis owing to their high similarity with ID0 ("nose"). Therefore, this method uses 13 keypoints, excluding those related to the face.

3.2.2 Task feature learning function (ST-GCN)

This function applies a model specialized for time-series analysis, such as ST-GCN, to construct a task classification model. ST-GCN is a graph neural network designed to handle unstructured data, such as skeletal data, and has the ability to learn both spatial and temporal features in an integrated manner. Furthermore, ST-GCN has a relatively simple model structure and a low inference load, making it suitable for practical use in multiple factories, as it offers low implementation costs and easy management. The input data consists of the coordinates of 13 keypoints based on the "nose" and their temporal variations, which are fed into the model. The skeletal information is modeled as a graph, with nodes (keypoints) and edges (relationships between nodes), and spatial and temporal convolutions are used to analyze in detail dynamic motion patterns such as walking and jumping. This approach is particularly effective for complex actions, such as those performed in factory work, and excels in capturing both the spatial configuration of the skeleton and its temporal changes simultaneously. In this method, the analysis results from ST-GCN are used to evaluate the model's characteristics and applicability.

3.2.3 Task classification function (ST-GCN)

This function utilizes the classification model constructed using the task feature learning function to perform task classification based on the input skeletal information. In this function, the evaluation data is extracted from a dataset different from the training data. ST-GCN is applied in this function, and its classification results are then evaluated.

Table 1
List of keypoints utilized in the system.

ID	0	5	6	7	8	9	10
Keypoint	Nose	Left shoulder	Right shoulder	Left elbow	Right elbow	Left wrist	Right wrist
	11	12	13	14	15	16	
	Left buttock	Right buttock	Left knee	Right knee	Left ankle	Right ankle	

3.3 Estimated task time

In this method, task classification is performed using the two-stage model shown in Fig. 3, and task time estimation is conducted on the basis of the classification results. First, the first-stage model in the framework is used to recognize the work objects; subsequently, the second-stage model classifies the specific task actions, enabling the identification of the start and end times for each task. This allows for the automatic estimation of the execution time for each individual task.

To evaluate the estimation accuracy of each task, we use the relative error rate based on the total task time. This metric is calculated by summing the estimated times for all instances within a task and comparing the result with the total measured time. The relative error rate is defined as

$$Relative\,Error = \left| \frac{T_{total}^{pred} - T_{total}^{true}}{T_{total}^{true}} \right|. \tag{1}$$

In this research, the accuracy of task time estimation for each task was evaluated by reporting the relative error as a percentage. By performing the evaluation based on the total time, we focus not on small errors in individual instances but on how accurately the total task time for the entire process is estimated.

Additionally, in this research, the mean and variance of the estimated task time for each task are calculated to assess the consistency (stability) of task time and the variability in task difficulty. The mean and variance of task time are evaluated using the following formulas:

$$T_{ave} = \frac{1}{N} \sum_{i=1}^{N} T_i^{pred}, \tag{2}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (T_i^{pred} - T_{avg})^2}\,. \tag{3}$$

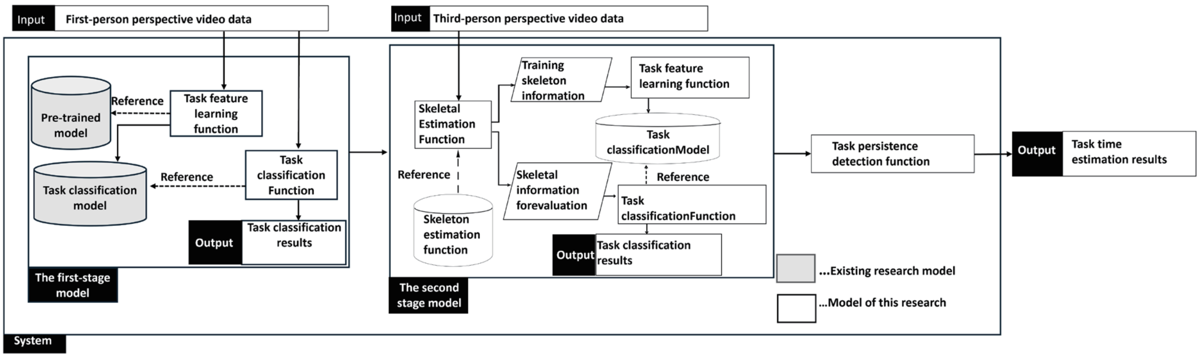
Input
First-person perspective video data
Pre-trained model
Reference
Task feature learning function
Task classification model
Reference
Task classification Function
Output
Task classification results
The first-stage model
Input
Third-person perspective video data
Skeletal Estimation Function
Reference
Skeleton estimation function
Training skeleton information
Task feature learning function
Task classificationModel
Skeletal information forevaluation
Reference
Task classificationFunction
Output
Task classification results
The second stage model
Task persistence detection function
Output
Task time estimation results
...Existing research model
...Model of this research
System

Fig. 3. System overview of estimated task time.

These metrics enable the quantitative evaluation of the stability and variability of time spent in each task.

3.4 Task persistence detection function

The condition for continuing the estimation of a task is set as follows: To determine when a task changes, an observation period of 6 s (180 frames) is used. If no change in the task occurs within this period, the task is considered to remain within the same task, and the task time estimation continues. This approach enables task time estimation that takes into account the continuity of the tasks.

The observation window was set to 6 s (180 frames), considering the physical characteristics of the dyeing process. Since typical tasks in this environment—such as handling large fabrics or operating machinery—generally persist for several seconds, a 6 s window provides sufficient temporal context. A shorter window would be overly sensitive to transient noise and minor fluctuations in movement, whereas an excessively long window would hinder the precise detection of task transitions. Therefore, 6 s was chosen as a reasonable duration to ensure stable classification while maintaining temporal sensitivity.

4. Experiment

The fundamental structure, learning conditions, and data acquisition procedures of the two-stage model employed in this research were previously documented in detail by the authors. For reproducibility, only the keypoints are briefly outlined here, and the reader is referred to the previous work[(4)] for further details.

However, a departure from previous research was necessary to perform task time estimation in this research. To this end, the training data for the second-stage model were newly collected, and the dataset was expanded.

4.1 Experimental method for analysis using a two-stage model

In this experimental procedure, task classification is performed using both first-person and third-person perspective video footage. The experimental steps involve performing the first-stage model experiment, followed by the second-stage model experiment. In the analysis using first-person perspective video, the initial model categorizes the data into five classifications, whereas the subsequent model categorizes the data into seven classifications. Specifically, items classified as “Sewing machine” in the first-stage model are further categorized as “Sewing machine operation” and “Sewing machine preparation” in the second-stage model. Items classified as “Dyeing machine” are further subdivided into the following categories: “Dyeing machine lid operation”, “Fabric cutting”, and “Panel operation”. Items classified as “Dyeing machine and cart” are further categorized into two distinct groupings: “Fabric insertion” and “Fabric removal”.

The dataset used in this experiment comprises images captured from the subject's viewpoint and videos documenting the dyeing process, recorded in a textile processing factory. The number of data instances and task categories for the second-stage model are presented in Table 2.

4.2 Analysis methods for estimating task time

In this research, task time estimation is performed using both first-person and third-person perspective video footage. The experimental procedure involves first using the two-stage model to classify the tasks, and then performing task time estimation on the basis of the classification results. By automatically identifying the start and end times of each task, the execution time for each task is calculated. The estimated task times are then compared with the actual task times, and the accuracy is evaluated to assess the estimation capability. The task categories used are the same as those employed in the second-stage model.

The condition for continuing the estimation of a task is defined as follows: to determine when a task changes, a 6 s (180-frame) observation period is used. If no change in the task occurs within this period, the task is considered to be part of the same task, and the time estimation continues. This approach allows for task time estimation that takes task continuity into account.

5. Experimental Results and Discussion

5.1 Two-stage model

The experimental results for the first-stage model were obtained using the same model configuration and learning conditions as in the authors' previous research, and the classification performance is consistent with that of the previous research. In this research, these results are used in preprocessing for the two-stage model and task time estimation.

The task classification results for the second-stage model in this research are shown in Tables 3–5, with a comparison with the previous classification results presented in Table 6. As shown in Table 6, an improvement in accuracy was observed in all categories. Notably, in categories with subtle differences between actions, such as "dyeing machine lid operation" and "fabric cutting", the misclassification rate significantly decreased. These categories were particularly challenging owing to unclear task boundaries, where minor posture changes or transitions between actions led to misclassification. However, by increasing the number of training instances to 1500 in each

Table 2
Dataset for analysis using third-person perspective video footage.

Task	Sewing machine operation	Sewing machine preparation	Dyeing machine lid operation	Fabric cutting	Panel operation	Fabric insertion	Fabric removal
Training data	1500	1500	1500	1500	1500	1500	1500
Test data	200	200	200	200	200	200	200

Table 3
Experimental results of analysis using third-person perspective video with two-stage model (sewing machine).

Task	Sewing machine operation	Sewing machine preparation	Average
Precision	0.967	0.996	0.981
Recall	0.995	0.968	0.982
F1 score	0.981	0.981	0.981

Table 4
Experimental results of analysis using third-person perspective video with two-stage model (dyeing machine and cart).

Task	Fabric insertion	Fabric removal	Average
Precision	0.938	0.996	0.967
Recall	0.995	0.941	0.968
F1 score	0.966	0.968	0.967

Table 5
Experimental results of analysis using third-person perspective video with two-stage model (dyeing machine).

Task	Dyeing machine Lid Operation	Fabric cutting	Panel operation	Average
Precision	0.900	0.900	1.000	0.933
Recall	0.918	0.957	0.926	0.934
F1 score	0.909	0.928	0.962	0.933

Table 6
Comparison with previous experimental results.

Task	Sewing machine operation	Sewing machine preparation	Fabric insertion	Fabric removal	Dyeing machine lid operation	Fabric cutting	Panel operation
Precision	0.059	0.168	0.040	0.181	0.080	0.310	0.032
Recall	0.224	0.071	0.127	0.107	0.156	0.140	0.113
F1 score	0.147	0.165	0.083	0.108	0.120	0.246	0.079

category, it is considered that ST-GCN was able to better capture the temporal and spatial features of the actions, resulting in a reduction in the frequency of misclassification.

Regarding the increase in the amount of training data for the second-stage model, there is generally a risk of overfitting if the additional data lack diversity relative to the model capacity. However, in this study, the added data are not simple replications collected under identical conditions. Instead, they encompass real-world samples captured across different workers, times of day, and working situations, including recordings from different dates. This enhances the diversity of the input distribution, thereby improving the model's generalization ability rather than promoting overfitting. Furthermore, no significant performance gap was observed between the training and test sets during the experiments, suggesting that severe overfitting did not occur. Although the current evaluation is limited to a specific factory environment, future work will involve testing the system under broader conditions, such as different equipment layouts and lighting environments, to further validate its robustness.

5.2 Work time estimation

In Table 7, the task time estimation results before the expansion of the training data for the second-stage model are shown, whereas Table 8 shows the results after data expansion. In both cases, the relative error for the estimated time is calculated for each task.

First, we examined the results obtained prior to increasing the number of training instances for the second-stage model. We found that the relative error for the total estimated time of each task ranged from approximately 12.5 to 79.0%. This significant variation in estimation accuracy depended on the type of task. Notably, in the categories “sewing machine operation”, “sewing machine preparation”, and “dyeing machine lid operation”, we found that the relative errors exceeded 50%, which was substantially higher than those observed in other categories (which ranged from 12.5 to 27.0%). These results highlighted the specific technical limitations of the initial model when applied to tasks with complex motion patterns. The tasks “sewing machine operation” and “sewing machine preparation” often have unclear boundaries in real-world scenarios, and during the transition from preparation to operation, similar postures and movements appear consecutively. As a result, the action label transition in the second-stage model tends to fluctuate, causing the misalignment of the start and end frames, which leads to larger relative errors in time estimation. For “dyeing machine lid operation”, the motion of opening and closing the lid often resembles the actions of fabric cutting or adjustments made by hand, making it difficult to distinguish these actions solely on the basis of skeletal information, and continuous misclassification likely contributed to the increased error.

In contrast, after increasing the amount of training data for the second-stage model, the results show a trend of decreased relative error across all categories. Specifically, for the previously mentioned categories—”sewing machine operation”, “sewing machine preparation”, and “dyeing machine lid operation”—a significant reduction in relative error was observed, clearly indicating the effectiveness of data augmentation. Additional data allowed for a more

Table 7
Task time estimation results (before data augmentation).

Task	Fabric insertion	Fabric removal	Sewing machine operation	Sewing machine setup	Dyeing machine lid operation	Fabric cutting	Panel operation
Predicted time (s)	4640.00	3858.00	2772.50	309.00	202.00	1330.50	2295.00
Actual time (s)	3697.50	4485.00	1830.00	1057.50	960.00	1822.50	2040.00
Relative error (%)	25.49	13.98	51.49	70.76	78.95	26.98	12.50

Table 8
Task time estimation results (after data augmentation).

Task	Fabric insertion	Fabric removal	Sewing machine operation	Sewing machine setup	Dyeing machine lid operation	Fabric cutting	Panel operation
Predicted time (s)	4090.50	4117.50	2010.00	877.50	832.50	1830.00	2160.00
Actual time(s)	3697.50	4485.00	1830.00	1057.50	960.00	1822.50	2040.00
Relative error (%)	8.30	6.90	9.80	17.50	13.00	0.40	5.90

comprehensive coverage of posture changes and transition patterns, enabling the second-stage model to classify frames near task boundaries more accurately, thereby reducing the occurrence of continuous misclassifications.

Moreover, even for categories that originally had small relative errors, the overall errors were further reduced, suggesting that the increase in data volume not only benefited specific categories but also contributed to improving the overall stability of the model. These results suggested that the increase in the number of training instances enhanced the model's ability to generalize across diverse motion patterns and worker postures. This improved generalization reduced local fluctuations in the time-series labels, making the estimation of start and end frames more consistent. Overall, expanding the training data for the second-stage model led to a reduction in relative error in task time estimation, confirming that data augmentation contributes to the improvement of model performance.

Table 9 shows the average and standard deviation of estimated task times for each task. The tasks "fabric removal" and "fabric insertion" have the longest average times, indicating that these are significantly longer processes than the other tasks. On the other hand, "panel operation" has the shortest average time (13.4 s), and tasks such as "dyeing machine lid operation", "fabric cutting", and "sewing machine preparation" are relatively shorter.

When focusing on the standard deviation, "panel operation" and "sewing machine preparation" have small variations, suggesting that these tasks are more standardized and performed consistently. In contrast, tasks such as "fabric insertion", "fabric removal", "sewing machine operation", and "fabric cutting" show larger standard deviations, indicating that there is greater variation in the execution times of these tasks. Although "dyeing machine lid operation" has a short average time, its standard deviation is 13.1 s, indicating a relatively high variability in time compared with other short-duration tasks.

The tasks "fabric insertion" and "fabric removal" are considered to be dominant bottlenecks for the overall cycle time of the process, as they have both long average times and large standard deviations. These tasks are highly affected by environmental conditions such as the worker's walking distance, the amount of fabric, and how the fabric is handled, which likely results in significant time variation.

On the other hand, "panel operation" and "sewing machine preparation" have both short average times and small standard deviations, suggesting that these tasks are relatively well standardized, with small differences between workers. These tasks are already stable, and there is currently less need for improvement.

For "sewing machine operation" and "fabric cutting", the average time is moderate, but the standard deviation is large, indicating that these tasks have variations depending on the worker or the situation.

Table 9
Mean task time and standard deviation.

Task	Fabric insertion	Fabric removal	Sewing machine operation	Sewing machine setup	Dyeing machine lid operation	Fabric cutting	Panel operation
Mean time (s)	145.2	158.6	45.3	32.5	22.7	28.9	13.4
Standard deviation	28.0	24.9	17.5	9.7	13.1	17.5	6.0

5.3 Comparison with existing methods

High action recognition accuracy was achieved in previous work by focusing on hand–object interactions.[8,11] However, in dyeing processes, hands and manipulated objects are frequently obscured by large machinery and materials, which can degrade the performance of interaction-dependent approaches. To address this, our two-stage model leverages both object context from the egocentric view and skeletal dynamics from the third-person view, allowing the system to compensate for unreliable local cues. This design ensures robustness in real-world industrial environments where occlusions are inevitable.

Moreover, while some existing approaches offer superior accuracy under controlled conditions, they often necessitate high-fidelity sensors and stringent setup requirements, leading to increased operational costs. By contrast, our system utilizes off-the-shelf cameras, prioritizing cost-effectiveness and ease of deployment. Our contribution lies in providing a practical, deployable solution that maintains reliable performance for task time estimation even amidst the frequent occlusions characteristic of dyeing plants.

5. Conclusions

We established in this work an effective sensing framework for automated task-time estimation. By utilizing cameras as noncontact sensors, the proposed system provides a practical and cost-effective sensing solution for digitizing traditional manufacturing processes.

In this research, a two-stage model combining first-person and third-person perspective video-based skeletal information was used to automatically classify tasks and estimate task times in the dyeing process. The task object classification using VGG16 and the task action classification using ST-GCN, as proposed in previous research, were reused, while the training data for the second-stage model was expanded. Additionally, a task time estimation method based on the output label sequence of the two-stage model was developed.

As a result, we found that increasing the number of training instances to 1500 per category for the second-stage model led to marked reductions in classification and relative errors in task time estimation across all categories. Notably, tasks such as “sewing machine operation”, “sewing machine preparation”, and “dyeing machine lid operation”, which tend to have unclear boundaries and are prone to misclassification, showed a significant improvement in relative error. This demonstrates that data augmentation for the skeletal time-series model is effective not only for improving action classification but also for enhancing the accuracy of task time estimation.

Furthermore, the average and standard deviation of estimated task times for each task were calculated to analyze the time characteristics of each process. The results revealed that “fabric insertion” and “fabric removal”, with long average times and large standard deviations, could be the bottlenecks governing the overall cycle time of the process. In contrast, tasks such as “panel operation” and “sewing machine preparation”, with short average times and small standard deviations, were considered relatively standardized and stable processes. Thus, the task time information obtained through the proposed method serves as a useful indicator for prioritizing process improvement and task standardization.

The significance of this research lies in presenting a framework that integrates action classification and task time analysis using only camera-based measurements, without disrupting the work flow on-site. This method can be regarded as a foundational technology for advancing human-centric, data-driven task analysis in manufacturing environments, where labor shortages and skill transmission are critical issues.

Although this research was validated on a dyeing process, the proposed two-stage framework is not inherently dyeing-specific. It is expected to be applicable to other manufacturing processes where tasks are reasonably constrained by the manipulated objects and worker motions can be captured by a fixed third-person camera; however, the target objects, task categories, and training data must be redesigned and collected for each process. However, the research was limited to a single factory, and further data collection and model validation are required to assess the applicability of this approach to other factories.

As model compression and lightweight architectures continue to improve, future work will focus on investigating further the trade-off between accuracy and computational efficiency. We plan to evaluate newer lightweight backbones and compressed variants to ensure that the system is highly suitable for low-cost edge deployment in factory settings.

Acknowledgments

The authors would like to extend their sincere appreciation to KOMATSU MATERE Co., Ltd. for their generous cooperation and support throughout the course of this research.

References

1 L. Permata, and S Hartanti: Int. J. Manag. Appl. Sci. **2** (2016) 10. http://ijmas.iraj.in/paper_detail.php?paper_id=6148&name=Work_Measurement_Approach_to_Determine_Standard_Time_in_Assembly_Line

2 B. Hasanain: Machines **12** (2025) 3. https://doi.org/10.3390/machines12030159

3 F. B. Gilbreth: Motion Study: A Method for Increasing the Efficiency of the Workman (D. Van Nostrand Company, New York, 1921) Chap. 1.

4 F. Shibata, J. Kikuti, A. Mimura, I. Iwata, Y. Ueda, and K. Sakamoto: JSCE J. Appl. Mech. Struct. Eng. **6** (2025) 1. https://doi.org/10.11532/jsceiii.6.1_384

5 K. Simonyan and A. Zisserman: Proc. Int. Conf. Learning Representations (ICLR) (2015). https://doi.org/10.48550/arXiv.1409.1556

6 OpenMMLab Contributors: OpenMMLab Pose Estimation Toolbox and Benchmark: https://github.com/open-mmlab/mmpose (accessed December 2024).

7 S. Yan, Y. Xiong, and D. Lin: Proc. 32nd AAAI Conf. Artificial Intelligence (AAAI Press, 2018) 7444−7452. https://doi.org/10.48550/arXiv.1801.07455

8 T. Nakano and K. Shida: Comput. Ind. Eng. **74** (2023) 90. https://doi.org/10.11221/jima.74.90

9 H. Yoshikawa, S. Kaneko, T. Urano, H. Nagayoshi, and T. Ohta: Hitachi Hyoron **70** (2021) 2. https://www.hitachihyoron.com/rev/archive/2021/r2021_02/pdf/02b03.pdf

10 K. Murai, T. Imai, K. Arai, and T. Kobayashi: IPSJ Trans. **10** (2020) 3. https://ipsj.ixsq.nii.ac.jp/records/207248

11 A. Matin, M. R. Islam, Y. Zhu, and X. Wang: Int. J. Comput. Vis. Signal Process. **14** (2024) 9. https://www.semanticscholar.org/paper/Hybrid-Deep-Learning-for-Assembly-Action-in-Smart-Matin-Islam/a0fd24d350aa363e54703c91b17b17c1b4af17de

12 X. Li, Y. Zhang, J. Zhang, Y. Chen, S. Chen, Y. Gu, M. Zhou, R. A. Farneth, I. Marsic, and R. S. Burd: Proc. Estim. Phase Detect. Seq. Process. (2017). https://arxiv.org/abs/1702.08623

13 E. H. Grosse, F. Sgarbossa, C. Berlin, and W. P. Neumann: Int. J. Proc. Res. **61** (2023) 7749. https://doi.org/10.1080/00207543.2023.2246783

14 Y. Wanyan, X. Yang, W. Dong, and C. Xu: Int. J. Comput. Vis. **133** (2025) 6832. https://doi.org/10.1007/s11263-025-02503-6

15 J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2009) 248–255. https://doi.org/10.1109/CVPR.2009.5206848

About the Authors

Kazuma Sakamoto received his B.S., M.S., and Ph.D. degrees in informatics from Kansai University, Japan, in 2015, 2018, and 2021, respectively. Since 2021, he has been an assistant professor at the Faculty of Production Systems Engineering and Sciences, Komatsu University, Japan. His research interests include web mining, natural language processing, image processing, and sports informatics. (kazuma.sakamoto@komatsu-u.ac.jp)

Fuya Shibata received his B.S. degree in production systems engineering and sciences from Komatsu University, Japan, in 2024, where he is currently pursuing his M.S. degree at the Graduate School of Sustainable Systems Science. His research interests include video analysis.
(24211010@komatsu-u.ac.jp)

Iori Iwata received his M.S. degree from the Graduate School of Sustainable Systems Science of Komatsu University, Japan, in 2024, where he is currently pursuing his Ph.D. degree. His research interests include image recognition, tactical analysis, and sports informatics system development.
(24311001@komatsu-u.ac.jp)

Aki Mimura received his master's degree from the Graduate School of Sustainable Systems Science of Komatsu University, Japan, in 2024 and is currently employed at KOMATSU MATERE Co., Ltd. His research areas include behavior recognition and system development.
(ak_mimura@komatsumatere.co.jp)

Yoshihiro Ueda received his Ph.D. degree in mathematical information science from Kanazawa University, Japan, in 2001. Since 2021, he has been a professor at the Faculty of Production Systems Engineering and Sciences, Komatsu University, Japan. His research interests include productivity improvement, labor saving, automation using data science, and human centric system development. (yoshihiro.ueda@komatsu-u.ac.jp)