Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue



Cite this: *Digital Discovery*, 2024, **3**, 2458

# Application of object detection and action recognition toward automated recognition of chemical experiments†

Ryosuke Sasaki,[a] Mikito Fujinami[b] and Hiromi Nakai *[ab]

Received 21st January 2024
Accepted 15th October 2024

DOI: 10.1039/d4dd00015c

rsc.li/digitaldiscovery

Developments in deep learning-based computer vision technology have significantly improved the performance of applied research. The use of image recognition methods to manually conduct chemical experiments is promising for digitizing traditional practices in terms of experimental recording, hazard management, and educational applications. This study investigated the feasibility of automatically recognizing manual chemical experiments using recent image recognition technology. Both object detection and action recognition were evaluated, that is, the identification of the locations and types of objects in images and the inference of human actions in videos. The image and video datasets for the chemical experiments were originally constructed by capturing scenes from actual organic chemistry laboratories. The assessment of inference accuracy indicates that image recognition methods can effectively detect chemical apparatuses and classify manipulations in experiments.



## 1. Introduction

Over the past decade, artificial intelligence (AI) technologies, particularly machine learning (ML) and deep learning, have made tremendous progress and are now used in practical applications. Computer vision has emerged as one of the most successful applications of deep learning, particularly in the areas of image and video recognition. For instance, object detection, which involves identifying and locating target objects within an image and classifying them, has been effectively utilized across numerous fields such as manufacturing, robotics, healthcare, security systems, traffic monitoring, agriculture, and environmental management.[1,2] Similarly, action recognition, which assigns predefined labels to human and object movements in videos, is applied in video retrieval, visual surveillance, human–robot interaction, and autonomous driving.[3,4]

Chemical experiments that traditionally rely on manual procedures would benefit from the application of AI technologies for convenience. For example, recent studies have focused on creating datasets for chemical experiments aimed at object detection[5,6] and segmentation, evaluating their recognition accuracy[7] using deep learning-based methods. Another research has explored augmenting image data of chemical apparatuses through the artificial combination of diverse images to enhance object detection.[8] However, these efforts have predominantly concentrated on identifying chemical objects. Understanding manual chemical experiments goes beyond object identification and requires recognizing the manipulative actions of the experimenter. To the best of our knowledge, there have been no reports on applying action recognition techniques to chemical experiments. This gap highlights an opportunity for further research in the application of AI for comprehensive analysis and automation in chemical experimentations.

This study aims to automatically recognize chemical experiments using image recognition technology. Combining the information obtained from object detection and action recognition is expected to be a promising approach for automatically recognizing chemical experiments. Previously, we constructed an image dataset of chemical apparatuses for object detection.[9,10] This study presents the performance of object detection using an image dataset. In addition, a video dataset was constructed and applied to an action recognition method. The assessment will demonstrate the proof of concepts to explore the feasibility of utilizing action recognition in chemical experiments manipulation.

The structure of this article is as follows: Section 2 provides details on the adopted chemical datasets, with a specific focus on the video dataset. Section 3 describes the applied image recognition techniques and computational details. Section 4 presents the performance results of applying the image recognition methods to the datasets. Finally, concluding remarks are presented in Section 5.

[a]*Department of Chemistry and Biochemistry, School of Advanced Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan. E-mail: nakai@waseda.jp*

[b]*Waseda Research Institute for Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00015c

## 2. Data description

This section describes the datasets used to construct the ML models. Image and video datasets were originally constructed for object detection and action recognition. The image dataset adopted in this study has been reported in detail.[9,10] Briefly, an overview of the dataset is provided. The images were captured from videos recorded in organic chemistry laboratories, resulting in 5078 images. Seven objects were adopted for annotation, including the experimenter's hand and six common chemical apparatuses: a conical beaker, an eggplant-shaped flask, an Erlenmeyer flask, a pipette, a reagent bottle, and a separatory funnel. The annotations were manually provided in the You Only Look Once (YOLO)[11] dataset format using the Visual Object Tagging Tool (VoTT).[12] The image dataset is divided into training, validation, and test subsets consisting of 4304 images with 12 041 objects, 570 images with 1775 objects, and 205 images with 405 objects, respectively. The complete information, including the number of objects in each subset, is shown in data article.[9]

A chemical experiment video dataset was constructed for action recognition. The videos were recorded in organic chemistry laboratories using a fixed camera. Some videos were provided as e-learning materials for a chemical experiment laboratory class in the faculty. Approximately 10 s clips were selected from the videos, and action labels were assigned manually. The video dataset format aligns with UCF-101,[13] a standard dataset for action recognition.

Three actions were selected to complement object detection and to understand chemical manipulations: "adding," "stirring," and "transferring." "Adding" involves manipulations such as adding reagents or solutions between apparatuses, using pipettes, or dispensing spoons. "Stirring" included stirring with a glass rod, shaking, or inverting the apparatus. "Transferring" refers to the manual movement of apparatuses and reagent bottles. Fig. 1 illustrates representative examples of these actions from four frames clipped from the videos. Fig. 1(a) shows the "adding" sample, which is the transfer of a solution from the conical beaker to the Erlenmeyer flask. Fig. 1(b) represents "stirring," which involves holding the top of the eggplant-shaped flask and stirring. Fig. 1(c) is the "transferring" sample, which involves grabbing the top of the conical beaker and moving it from one end of the screen to the other. Three video samples for each action, including the representative examples shown in Fig. 1, are provided in the ESI.†

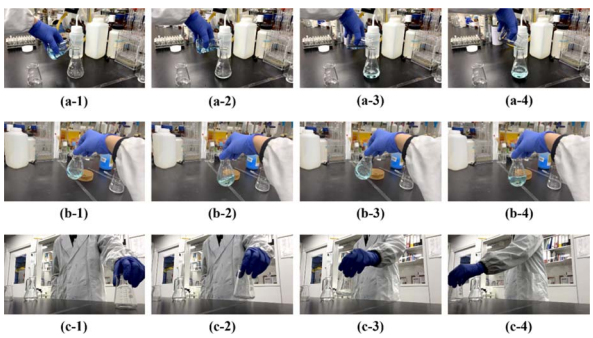


Fig. 1 Examples of chemical experiment video dataset for action recognition. (a)–(c) Represent the sample videos of "adding", "stirring", and "transferring", respectively. The images display four frames clipped from each video.

Table 1 The details of the constructed video dataset. The number of videos for three types of actions: "adding", "stirring", and "transferring" in training, validation, and test datasets are listed

| Class | Training | Validation | Test |
|---|---|---|---|
| Adding | 158 | 53 | 18 |
| Stirring | 91 | 31 | 18 |
| Transferring | 72 | 25 | 12 |
| Total | 321 | 109 | 48 |

The video dataset was divided into training, validation, and testing subsets. Table 1 lists the number of videos and corresponding filmed actions across the subsets. A total of 478 videos were created and divided into 321 videos for training, 109 videos for validation, and 48 videos for testing. For both image and video datasets, the training, validation, and test subsets were extracted from independent, non-overlapping videos to prevent data leakage between the subsets. Although the images and videos contain the same type of apparatuses and laboratory viewing in the subsets, the dataset includes diverse backgrounds and situations surrounding the objects.

## 3. Computational methods

This section explains the object detection and action recognition methodologies. In the selection of image recognition methods prediction timing, accuracy, and ease of implementation were considered. Object detection estimates the locations and their classes of objects in an image. Two primary schemes are known for object detection: one-stage methods that simultaneously predict both object location and label, and two-stage methods that independently predict object regions and class labels. In general, one-stage methods excel in prediction speed, whereas two-stage methods demonstrate higher accuracy. In this study, we applied YOLOv8,[14] a leading one-stage model. Action recognition assigns human movements in videos to predefined labels. In this study, a 3D Residual Network (ResNet)[15,16] was employed as the action recognition method. 3D ResNet integrates the ResNet architecture into a 3D convolutional neural network to extract spatiotemporal features. The detailed hyperparameters are described in Appendix section A1.

## 4. Results & discussion

In this section, the inference performance of the image recognition methods, object detection followed by action recognition, is presented.

### 4.1. Object detection

The object detection was evaluated by utilizing the average precision (AP) and mean AP (mAP).[17] Detailed definition is

Table 2 Statistical evaluation, including mAP and APs, for the seven types of objects in the test data predictions obtained by YOLOv8n and YOLOv8x, is presented

| Model | $mAP_{50}$ | Hand | Conical beaker | Erlenmeyer flask | Reagent bottle | Pipette | Eggplant shaped flask | Separatory funnel |
|---|---|---|---|---|---|---|---|---|
| YOLOv8n | 0.855 | 0.732 | 0.951 | 0.978 | 0.880 | 0.532 | 0.933 | 0.981 |
| YOLOv8x | 0.890 | 0.749 | 0.972 | 0.984 | 0.943 | 0.625 | 0.962 | 0.995 |

described in Appendix section A2. Learning process is shown in Appendix A3 section. We confirmed that YOLOv8n and YOLOv8x were reliable and therefore utilized the models for the following assessments. Table 2 lists the AP and mAP of each class for the test data. The mAPs for the test data obtained using YOLOv8n and YOLOv8x were 0.855 and 0.890, respectively. These results suggested that object detection using the image dataset from the chemical experiment was effective. YOLOv8x made predictions on the test data with a higher mAP than YOLOv8n. The APs for the pipette and hand classes were lower than those for the other classes. The APs for the pipette class are 0.532 and 0.625 for YOLOv8n and YOLOv8x, respectively. For the hand class, the AP values were 0.732 and 0.749 for YOLOv8n and YOLOv8x, respectively. Conversely, the APs of the other classes exceeded 0.880 and 0.943 for YOLOv8n and YOLOv8x, respectively.

Fig. 2 shows examples of object detection for the test data obtained using YOLOv8x. The conical beaker was correctly recognized in Fig. 2(a). As shown in Fig. 2(b), two Erlenmeyer flasks containing solutions of different colors were accurately detected. As shown in Fig. 2(c), both the hand and pipette were correctly recognized. As shown in Fig. 2(d), the separatory funnel, four reagent bottles, and an eggplant-shaped flask were identified correctly. When the entire target object was captured at a relatively large size, a tendency toward accurate detection was observed. In addition, correct detection was achieved for multiple objects in an image, even when parts of the objects overlapped.

Fig. 3 shows examples of misrecognition and nondetection in object detection for the test data in YOLOv8x. As shown in Fig. 3(a), the hand covered with a white rubber glove was not detected. The Japanese skin-colored arm was misrecognized as a hand. As shown in Fig. 3(b), the hand is correctly recognized. The Erlenmeyer flasks held manually were not detected. In Fig. 3(c), the Erlenmeyer flask was correctly recognized. The pipette inserted into the flask was not detected. The Erlenmeyer flask on the left side was identified correctly, as shown in Fig. 3(d). The conical beaker on the right side was misrecognized as an Erlenmeyer flask. For hand detection, the misrecognized cases indicate that the object color is a critical factor in prediction. The difficulty in detection increases when the objects overlap each other. Changes in the rectangular area owing to the angle of the object also affect the recognition accuracy, particularly in the case of pipette detection. The pipette was enclosed in an elongated rectangle when captured horizontally or vertically, whereas it was enclosed in a large square when captured diagonally. The significant diversity in object color and area contributed to the lower prediction accuracy observed in the statistical evaluation of the hand and pipette. Increasing the variation in the training data would help mitigate color- and angle-based biases.

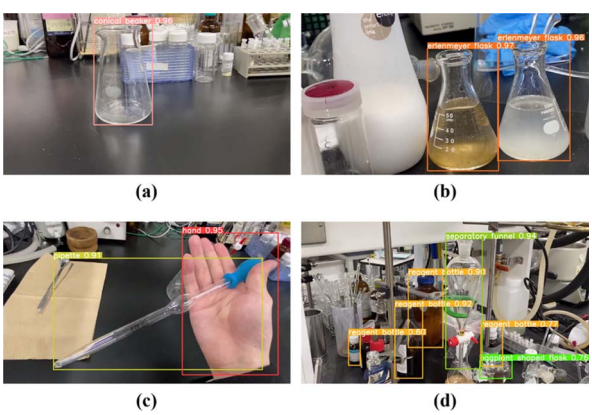


Fig. 2 Examples of object detection for test data using YOLOv8x. The detected objects are enclosed by rectangular frames with object labels.

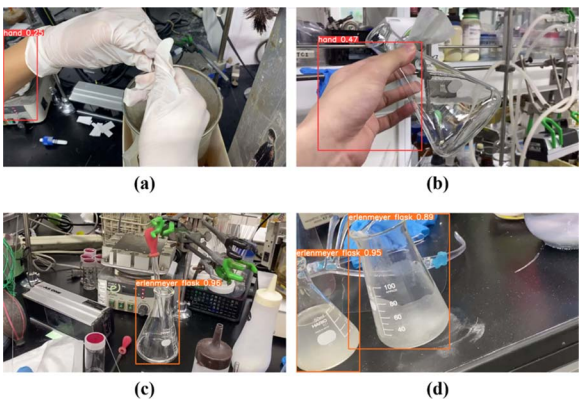


Fig. 3 Examples of misrecognition in object detection using YOLOv8x. The examples include cases where an object that should be recognized is not detected or is incorrectly labeled.

### 4.2. Action recognition

Action recognition is evaluated through accuracy, which quantifies the ratio of videos in which the predicted label with the highest probability matches the correct action. The learning process is shown in Appendix A3 section. Table 3 displays the prediction accuracy of the action recognition for the test data, listing the accuracy for the three actions and their average values. For the test data, classification accuracies for "adding," "stirring,"

Table 3 Statistical evaluation on action recognition. The prediction accuracy for the three types of actions and their average for the test datasets is presented

| Average | Adding | Stirring | Transferring |
|---|---|---|---|
| 0.86 | 0.94 | 0.89 | 0.74 |

and "transferring" were 0.94, 0.89, and 0.75, respectively, within an average accuracy of 0.86. These values suggest that the generalized action-recognition model was effectively trained.

Fig. 4 illustrates examples of action recognition for the test data. In Fig. 4(a), the "adding" of a solution from one conical beaker to another was correctly recognized. In Fig. 4(b), the "stirring" of the blue solution by a hand holding the top of the Erlenmeyer flask was correctly classified. In Fig. 4(c), the "transferring" of the conical beaker was correctly recognized. Fig. 5 shows examples of misrecognitions. In Fig. 5(a), the "stirring" of the blue solution in the beaker by hand was misidentified as "transferring." The confidence scores for classes, indicating the probability of the prediction being assigned to the corresponding action, were 0.61 and 0.35 for "transferring" and "stirring," respectively. In Fig. 5(b), the "transferring" of the Erlenmeyer flask by hand was misclassified as "stirring." The confidence scores were 0.70 and 0.30 for "stirring" and "transferring," respectively. In both misrecognition cases, "adding" exhibited a significantly low confidence score, indicating potential confusion between "transferring" and "stirring." Examples of misrecognition were observed in cases where hand and apparatus orientations were similar across different actions and where an action switched to another at the end of the video. These misrecognitions suggest that action recognition involves not only hand and object movements but also the type and angle of the object, and that a mixture of actions in a video has a negative effect on recognition. These findings emphasize the importance of data variation and meticulous data curation when constructing video datasets for action recognition. Notably, the assessment demonstrated an 86% accuracy

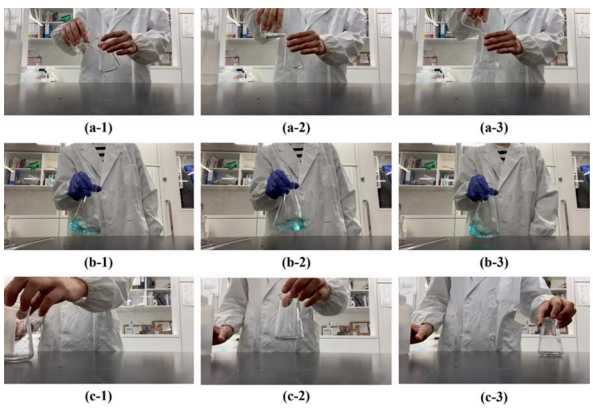


Fig. 4 Examples of action recognition for test data using 3D ResNet. (a)–(c) Depict the correctly classified actions of "adding," "stirring," and "transferring," respectively. The images display three frames clipped from each video.

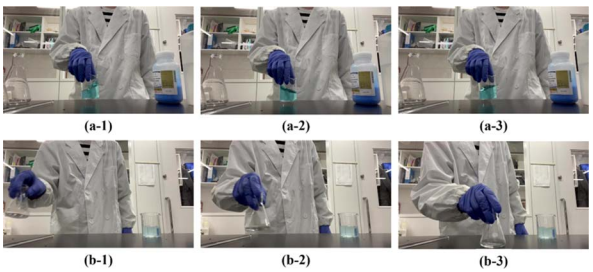


Fig. 5 Misrecognition examples of action recognition using 3D ResNet. In (a) and (b), "stirring" and "transferring" have been assigned reversely. The images display three frames clipped from each video.

prediction for the test data based on learning from 321 chemical experiment videos.

## 5. Concluding remarks

This study investigated the automatic recognition of chemical experiment images and videos using ML. Object detection and action recognition methods were applied to the constructed image and video datasets of the chemical experiments to assess their recognition capabilities. The image dataset comprises 5078 images, annotated for seven types of objects, whereas the video dataset consists of 478 videos, each featuring one of the three actions with corresponding labels. These datasets were manually curated using videos recorded in organic chemistry laboratories. Object detection achieved a recognition accuracy of 0.890 in mAP, whereas action recognition demonstrated a prediction accuracy of 86% for the test data. The prediction accuracy for both object detection and action recognition demonstrates that the trained models perform adequately when compared with benchmarks for common datasets using state-of-the-art methodologies. These results confirm that the application of image recognition methods to chemical experiment images and videos is effective.

The dataset constructed in this study is limited in both the size and the variety of labels, particularly in the case of video data. Although the manually filmed and curated datasets are highly reliable, the datasets lack diversity in terms of laboratory settings, personnel, and equipment. To evaluate the recognition accuracy on an entirely external dataset, the model trained in this study was applied to object detection on the LabPics dataset.[7] The detection accuracy was lower than the dataset used in this study. Detailed results of this verification are provided in the ESI.† To develop a universally applicable model across various experimental situations, a more extensive and diverse dataset is required. Developing a generally applicable model is anticipated to be a considerable challenge. As an alternative approach, building datasets and ML models that are specifically optimized for individual laboratories could be effective. In either case, developing a platform to partially automate data collection and model training may be a promising direction for future research.

Despite recent innovative developments such as the optimization of experimental conditions through high-throughput or flow reactors,[18–21] and the use of autonomous experiments facilitated by experimental chemical robots,[22–24] common

laboratories rely on manual procedures because of limitations in the applicable experimental protocols for specific automated equipment. The present image recognition of chemical experiment videos is anticipated to provide advantages for manual experiments, including automatic experiment recording, hazard warnings, and evaluation support for novice chemists, with minimal introduction costs and requiring only the installation of video cameras connected to the network.

## Data availability

The image datasets are publicly available in the following repository: **https://doi.org/10.17632/8p2hvgdvpn.1**. Representative examples of videos are provided in ESI.† Extra videos are available upon reasonable request to the authors. The machine learning methods utilized for this study are open source. The object detection method YOLOv8 is accessible at **https://github.com/ultralytics/ultralytics**. The action recognition method 3D ResNets can be found at **https://github.com/kenshohara/3D-ResNets-PyTorch**. The utilization in this paper and command example are shared at **https://github.com/rsasaki913/ChemImgRecog**.

## Author contributions

RS: methodology, software, formal analysis, investigation, visualization, data curation, writing – original draft. MF: validation, data curation, writing – original draft, funding acquisition. HN: conceptualization, validation, resources, writing – review & editing, supervision, project administration, funding acquisition.

## Conflicts of interest

There are no conflicts to declare.

## Appendices

### A1. Detailed parameters for machine learning

Both YOLOv8 (ref. 14) and 3D ResNet[16] were implemented using open-source programs based on PyTorch, a Python neural network library. The details of ML hyperparameters are as follows. YOLOv8 comprises five models, each with a varying number of model parameters. Numerical verification was conducted for YOLOv8n, which had the fewest parameters, and YOLOv8x, which had the most. The image size was set to 640 pixels, and the number of learning epochs was 300. Default settings were employed for the other hyperparameters. According to the default configuration of optimization algorithms, AdamW was applied for the first 38 epochs, followed by stochastic gradient descent optimization. Data augmentation, including adjustments to hue, saturation, and value (HSV), translation, scale, and mosaic, was applied using the default settings in the YOLOv8 implementation. The early termination occurred after 50 epochs, wherein learning was halted if the mAP against the validation data did not improve. $AP_{50–95}$ was used to evaluate the validation data and determine the optimal model, and the prediction accuracy of the test data was calculated using $AP_{50}$.

ResNet-34, a 3D ResNet model with 34 layers, was used for action recognition. Stochastic gradient descent optimization with momentum was applied using the weight decay and momentum values of 0.0005 and 0.9, respectively. The learning rate was set to 0.1 for the first 50 epochs, 0.01 from 51 to 100 epochs, and 0.001 from 101 to 5000 epochs. Data augmentation techniques, including four-corner/center cropping, horizontal flipping, and scaling of video clips, were employed using the 3D ResNet implementation. A pre-trained model that trained 200 epochs of Kinetics-700 (ref. 25) and Moments in Time[26] was adopted as the initial model for learning.

### A2. Evaluation metrics for object detection

The detailed definition of evaluation metrics for object detection, AP and mAP, is explained.[17] AP assesses the prediction accuracy for both object regions and classes, computed through the intersection over union (IoU), which represents the degree of overlap between two regions, as defined in eqn (1).

$$\mathrm{IoU} = \frac{A \cap B}{A \cup B} \tag{1}$$

Symbols $A$ and $B$ represent the regions where the predicted and correct objects exist, respectively. The predicted area where the IoU exceeded the threshold was utilized for classification. Precision and recall based on the classification results were applied to compute the AP, which is defined as the area under the precision–recall curve, ranging from zero to one, with higher values indicating better prediction accuracy. Two feature values are commonly used: namely, $AP_{50}$ for the IoU threshold set to 0.5 and $AP_{50–95}$ for the average AP obtained by varying the IoU threshold from 0.5 to 0.95, with a step size of 0.05.[27] Generally, the $AP_{50–95}$ provides a more stringent evaluation than the $AP_{50}$. The other parameter, mAP, is the average AP across all classes.

### A3. Learning process of object detection and action recognition

The optimal ML models of object detection and action recognition were determined using prediction accuracy for validation

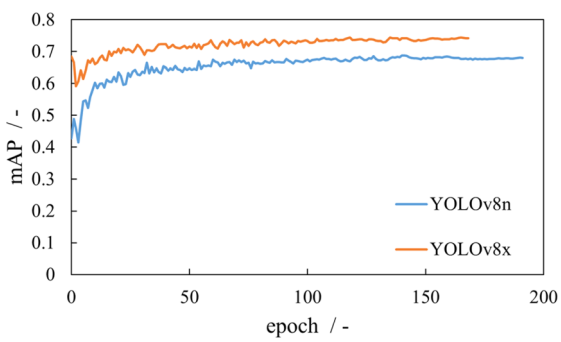


Fig. 6 Learning curve of object detection. The epoch and corresponding mAP for validation datasets during the training process of YOLOv8n and YOLOv8x are shown.

Table 4 Statistical evaluation, including mAP and APs, for the seven types of objects in the validation data predictions obtained by YOLOv8n and YOLOv8x, is presented

| Model | $mAP_{50}$ | Hand | Conical beaker | Erlenmeyer flask | Reagent bottle | Pipette | Eggplant shaped flask | Separatory funnel |
|---|---|---|---|---|---|---|---|---|
| YOLOv8n | 0.825 | 0.882 | 0.866 | 0.806 | 0.881 | 0.722 | 0.785 | 0.832 |
| YOLOv8x | 0.854 | 0.873 | 0.913 | 0.848 | 0.871 | 0.780 | 0.821 | 0.871 |

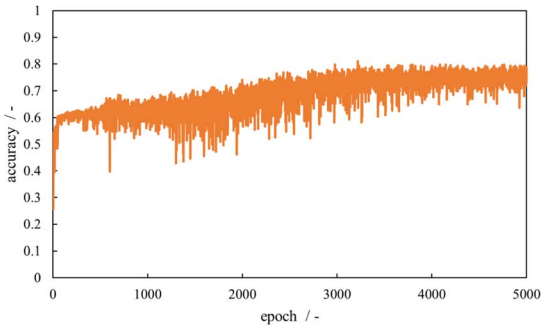


Fig. 7 Learning curve of action recognition using 3D ResNet. The epochs and corresponding accuracy for validation datasets are shown.

Table 5 Statistical evaluation on action recognition. The prediction accuracy for the three types of actions and their average for the validation datasets is presented

| Average | Adding | Stirring | Transferring |
|---|---|---|---|
| 0.80 | 0.96 | 0.84 | 0.60 |

data. The determined models were applied to test data for evaluating model performance. Fig. 6 illustrates the learning curve for YOLOv8n and YOLOv8x, where the horizontal axis represents the epoch and the vertical axis denotes the mAP. The blue and orange lines depict the mAP for the validation data using YOLOv8n and YOLOv8x, respectively. As the number of epochs increased, the validation mAPs improved, and convergence was observed in both YOLOv8n and YOLOv8x learning. YOLOv8n and YOLOv8x achieved the highest mAPs at 142 and 119 epochs, respectively. Table 4 provides the AP and mAP for each class in the validation data. The mAPs for the validation data obtained using YOLOv8n and YOLOv8x were 0.825 and 0.854, respectively. The learning curves and mAP values indicated that YOLOv8n and YOLOv8x were effectively trained.

Fig. 7 illustrates the learning curve for action recognition using 3D ResNet. The horizontal and vertical axes represent the epochs and prediction accuracy, respectively. The orange line indicates the accuracy of the validation data. The accuracy increased up to approximately 3000 epochs, demonstrating a tendency to converge with the oscillations. This behavior suggests that ML is progressing appropriately. The model with 3218 epochs displayed the highest prediction accuracy for the validation data. Table 5 shows the prediction accuracy for action recognition on the validation data, listing the accuracy for the three actions and their average values. For the validation data, the classification accuracies for "adding," "stirring," and "transferring" were 0.96, 0.84, and 0.60, respectively, with an average accuracy of 0.80. The model was selected as the optimal model and applied to the numerical verification.

## Acknowledgements

We thank Prof. Kanomata, Prof. Shibata, and their group members for their cooperation in capturing the images and videos of the chemical experiments. We appreciate the faculty laboratory staff for providing e-learning materials for the experimental videos. One of the authors (M. F.) is grateful for the financial support from The Grant-in-Aid for Transformative Research Areas (A) Digitalization-driven Transformative Organic Synthesis (Digi-TOS) (KAKENHI Grant Number JP22H05380) from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan.

## References

1 R. Kaur and S. Singh, *Digital Signal Processing*, 2023, **132**, 103812.
2 J. Terven, D.-M. C. Esparza and J.-A. R. González, *Machine Learning and Knowledge Extraction*, 2023, **5**, 1680–1716.
3 Y. Kong and Y. Fu, *International Journal of Computer Vision*, 2022, **130**, 1366–1401.
4 G. Diraco, G. Rescio, P. Siciliano and A. Leone, *Sensors*, 2023, **23**, 5281.
5 Z.-S. Ding, S.-Y. Ran, Z.-Z. Wu, Z.-H. He, Q.-Q. Chen, Y.-S. Wei, X.-F. Wang and L. Zou, A New Benchmark Data Set for Chemical Laboratory Apparatus Detection, in *Artificial Intelligence in Data and Big Data Processing Proceedings of ICABDE 2021*, 2022, pp. 201–210.
6 X. Cheng, S. Zhu, Z. Wang, C. Wang, X. Chen, Q. Zhu and L. Xie, *Artificial Intelligence Chemistry*, 2023, **1**, 100016.
7 S. Eppel, H. Xu, M. Bismuth and A. A. Guzik, *ACS Cent. Sci.*, 2022, **6**, 1743–1752.
8 S. Rostianingsih, A. Setiawan and C. I. Halim, *Procedia Computer Science*, 2020, **171**, 2445–2452.
9 R. Sasaki, M. Fujinami and H. Nakai, *Data Brief*, 2024, **52**, 110054.
10 Annotated Chemical Apparatus Image Dataset on Mendeley Data, **https://doi.org/10.17632/8p2hvgdvpn.1**, accessed June 2024.
11 J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

12 GitHub, **https://github.com/microsoft/VoTT**, accessed June 2024.
13 K. Soomro, A. R. Zamir and M. Shah, UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, *arXiv*, 2012, preprint, arXiv: 1212.0402, DOI: **10.48550/arXiv.1212.0402**.
14 GitHub, **https://github.com/ultralytics/ultralytics**, accessed June 2024.
15 H. Kataoka, T. Wakamiya, K. Hara and Y. Satoh, Would Mega-scale Datasets Further Enhance Spatiotemporal 3D CNNs?, *arXiv*, 2020, preprint, arXiv: 2004.04968, DOI: **10.48550/arXiv.2004.04968**.
16 GitHub, **https://github.com/kenshohara/3D-ResNets-PyTorch**, accessed June 2024.
17 R. Padilla, S. L. Netto and E. A. B. da Silva, A Survey on Performance Metrics for Object-Detection Algorithms, in *2020 International Conference on Systems, Signals and Image Proceeding (IWSSIP)*, 2020, pp. 237–242.
18 C. J. Taylor, A. Pomberger, K. C. Felton, R. Grainger, M. Barecka, T. W. Chamberlain, R. A. Bourne, C. N. Johnson and A. A. Lapkin, *Chem. Rev.*, 2023, **123**, 3089–3126.
19 L. Capaldo, Z. Wen and T. Noel, *Chem. Sci.*, 2023, **14**, 4230–4247.
20 C. W. Coley, N. S. Eyke and K. F. Jensen, *Angew. Chem., Int. Ed.*, 2020, **59**, 22858–22893.
21 L. Buglioni, F. Raymenants, A. Slattery, S. D. A. Zondag and T. Noel, *Chem. Rev.*, 2022, **122**, 2752–2906.
22 M. Seifrid, R. Pollice, A. Aguilar-Granda, Z. M. Chan, K. Hotta, C. T. Ser, J. Vestfrid, T. C. Wu and A. A. Guzik, *Acc. Chem. Res.*, 2022, **55**, 2454–2466.
23 L. Wilbraham, S. H. M. Mehr and L. Cronin, *Acc. Chem. Res.*, 2021, **54**, 253–262.
24 C. W. Coley, N. S. Eyke and K. F. Jensen, *Angew. Chem., Int. Ed.*, 2020, **59**, 23414–23436.
25 J. Carreira, E. Noland, C. Hiller and A. Zisserman, A Short Note on the Kinetics-700 Human Action Dataset, *arXiv*, 2022, preprint, arXiv: 1907.06987, DOI: **10.48550/arXiv.1907.06987**.
26 M. Monfort, A. Andonian, B. Zhou, K. Ramakrishman, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick and A. Oliva, Moments in Time Dataset: one million videos for event understanding, *arXiv*, 2019, preprint, arXiv: 1801.03150, DOI: **10.48550/arXiv.1801.03150**.
27 M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn and A. Zisserman, *International Journal of Computer Vision*, 2010, **88**, 303–338.