Affective and Pleasurable Design, Vol. 41, 2022, 129–137
https://doi.org/10.54941/ahfe1001789

Time-Series Segmentation Based on Video Images of Cutting Operations with a Lathe in Virtual Reality Space

Shohei Tawata[1], Keiichi Watanuki[1,2], and Kazunori Kaede[1,2]

[1]Graduate School of Science and Engineering, Saitama University, Shimo-Okubo 255, Sakura-ku, Saitama-shi, Saitama 3388570, Japan

[2]Advanced Institute of Innovative Technology, Shimo-Okubo 255, Sakura-ku, Saitama-shi, Saitama 3388570, Japan

ABSTRACT

Due to the shortage of labor force in Japan, skill transfer and training education are becoming increasingly important in the manufacturing industry. In recent years, virtual reality (VR) technology has attracted attention in work-related training, enabling simplified training, but there is a problem that human and time resources cannot be sufficiently allocated to training due to a lack of educators and an immature training system. In this study, we developed a method to automatically recognize tasks and actions to improve the efficiency of education and training. To recognize tasks and actions, we adopted a deep learning model that can recognize actions from videos in time series, and we pre-trained the model on a large open-source dataset. We evaluated the performance of the model on unlearned procedures and people by preparing a dataset with three different procedures and 10 participants. The overall validation metrics all exceeded 90%. Specifically, results of more than 90% were achieved for unlearned people, but a drop of more than 5% was observed for all unlearned procedures, suggesting that issues must be addressed for application to task training.

Keywords: Motion analysis, Human action segmentation, Work training, Virtual reality, Lathe

INTRODUCTION

Due to the shortage of labor force in Japan, skill transfer and training education are becoming increasingly important in the manufacturing industry. According to a 2021 Ministry of Economy, Trade and Industry (METI) survey, more than 40% of manufacturing companies cited a lack of progress in human resource and skills development as a management issue (METI, 2021). Insufficient training not only prevents work efficiency improvement but also increases the risk of work errors and omissions. Recently, the use of virtual reality (VR) technology has attracted attention as a way to conduct efficient training, with the advantages of smaller facilities than real-world environments, the ability to safely conduct training for dangerous tasks, and the ability to conduct repetitive training. While this makes it possible to conduct simple training, there is a problem of inadequate human and time resources due to a lack of educators and an immature training system, which

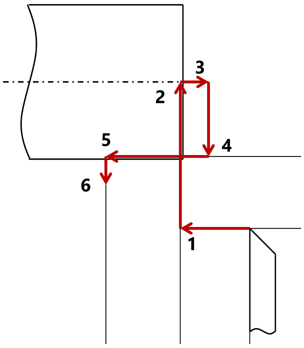

Figure 1: Reference task (Task 1).

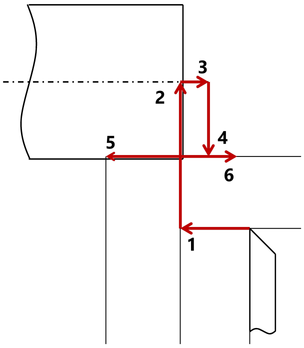

Figure 2: Task that assumes errors in work (Task 2).

prevents adequate feedback for trainees. Therefore, there is a need to improve the efficiency of training and education so that trainees can acquire sufficient skills even when there is a shortage of educators. Motion analysis, which divides a task into detailed actions, is an effective method for recognizing the overburdening, waste, and irregularity of a task. It is also useful for evaluating the performance of tasks during training. However, in the past, analysis was generally conducted by humans, which required significant time and effort, making it difficult to conduct detailed analysis for individual training sessions.

Therefore, in this study, we developed a method to automatically recognize tasks and actions to support the creation of educational content and self-learning for trainees. The target work was cutting using a lathe, which is practiced in mechanical engineering departments of universities, technical colleges, and technical high schools. An RGB camera was used because it enables comparison between the actual work and training. Several papers (Lea et al. 2017; Farha and Gall, 2019; Ishikawa et al. 2021) have proposed models that can perform action segmentation using RGB camera information. There are also models (Carreira and Zisserman, 2017) that can obtain feature extractors for videos by pre-training on large open-source datasets, which is expected to enable motion recognition without a large amount of labeling. In this study, we adopted one of the action segmentation models using a feature extractor that can be pre-trained, which has been

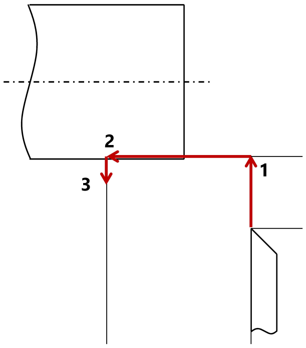

Figure 3: Task that assumes omissions in work (Task 3).

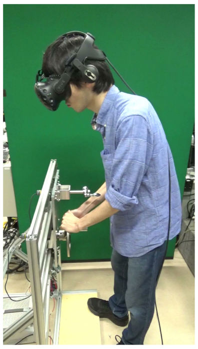

Figure 4: Example of VR operation scene.

very successful in recent years, and applied it to a video of a task using a VR lathe. Three types of tasks were prepared: a reference task (Task 1), task that assumes work errors (Task 2), and task that assumes work omissions (Task 3). The predictive ability of the model was evaluated for procedures and operator data that were not used when the model was trained.

VIDEO DATASET OF A CUTTING OPERATION IN VR SPACE

In this study, we prepared a video dataset of facing and outer round turning on a lathe system in VR space. A lathe was placed in the VR space, and the spindle was constantly rotating with a cylindrical material attached to the chuck. The carriage feed and lateral feed handles of the lathe could operate the carriage feed and lateral feed of the lateral feed table, respectively. Each of these handles returned a reaction force by a motor, and a larger reaction force was generated while contacting the work compared to not contacting it.

Figures 1 to 3 show the movements of the bite edge in the task. In Task 1, the work was cut in the order of facing and outer round turning. In Task 2, the direction of bite release was different to that in Task 1, representing an error in the direction of bite movement. In Task 3, only the outer round cutting from Task 1 was performed, representing a lapse of procedure. In each task, the cutting depth and feed rate were specified, and images showing the movement of the cutting edge were presented in the VR space so that the operator could check the contents of the task and current cutting depth and

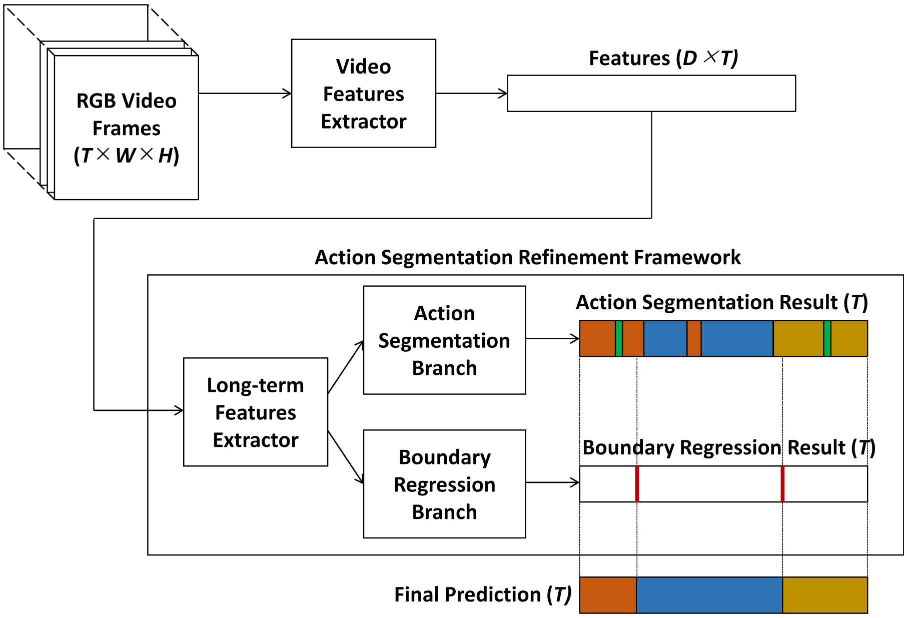
RGB Video Frames ($T \times W \times H$)
Video Features Extractor
Features ($D \times T$)
Action Segmentation Refinement Framework
Long-term Features Extractor
Action Segmentation Branch
Boundary Regression Branch
Action Segmentation Result (T)
Boundary Regression Result (T)
Final Prediction (T)

Figure 5: Overall structure of action segmentation refinement framework (ASRF).

feed rate at all times. Figure 4 shows an example captured image; the image was taken with the carriage feed and cross feed handles visible.

Under the above conditions, 10 participants took part in the experiment, and each participant filmed 10 times for Task and once for each of Tasks 2 and 3. Each frame was given one of six labels: Stop, Left (feeding left), Right (feeding right), Forward (cross feeding forward), Back (cross feeding backward), and Empty (transporting empty). The frame rate was 15 fps, and the resolution was resized to 240 pixels on the short side.

BUILDING THE ACTION SEGMENTATION MODEL

Model Structure

In this study, inflated 3D ConvNets (I3D) (Carreira and Zisserman, 2017) was used to extract features from video images, and the action segmentation refinement framework (ASRF) proposed by Ishikawa et al. (2021) was used as a learning model to output classification results. The system input was RGB video data $X = [\mathbf{x}_1, \cdots, \mathbf{x}_T] \in \mathbb{R}^{T \times C \times H \times W}$, where T is the number of frames of the input video, C is the number of channels ($C = 3$ in the case of RGB images), and H and W are the numbers of pixels in the height and width directions, respectively. X is the input to the processor and also the input to I3D, and the features output by I3D are represented as $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_T] \in \mathbb{R}^{T \times D}$. The features output by I3D are the input to ASRF, where D is the dimension of the features. The final output of ASRF is the classification result $\widehat{Z} = [\widehat{\mathbf{z}}_1, \cdots, \widehat{\mathbf{z}}_T] \in \mathbb{R}^{T \times K}$ and the probability of boundary $\widehat{B} = \left[\widehat{\mathbf{b}}_1, \cdots, \widehat{\mathbf{b}}_T\right] \in [0,\ 1]^T$, where K is the total number of classes. The output of the whole system is the frame-by-frame classification result $Z = [\mathbf{z}_1, \cdots, \mathbf{z}_T] \in \{0, 1\}^{T \times K}$ and the boundary $B = [\mathbf{b}_1, \cdots, \mathbf{b}_T] \in \{0, 1\}^T$. Figure 5 shows the overall structure of the model from input to output. We used the long-term features extractor (LEF), action segmentation branch (ASB), and boundary regression branch (BRB) including the temporal convolutional network (TCN) (Farha and Gall, 2019) with dilated residual layer (DRL); this model has been shown to be capable of action segmentation from time-series features, able to consider long-term dependencies, and highly robust to temporal resolution. In addition, the classification performance can be improved by overlapping the network into a multi-stage structure.

Model Training

In this study, the I3D model, which was pre-trained using the open-source datasets ImageNet and Kinetics 400, was used as a feature extractor, and the features extracted from the dataset of cutting operations in VR space were used as input to train the ASRF model, which is an action recognition model. The results are listed in Table 1. For training, 10 data for each of the 9 participants in Task (90 data in total) were used for training. The test data set consisted of one person's data (10 data) for Task 1 and one person's data (10 data) for each of Tasks 2 and 3. Due to the small size of the dataset, we conducted a 10-fold cross-validation, one person's data were used as the test data for Task 1 when evaluating the model. The Adam optimization algorithm was used with learning rate 0.0005 and batch size 1. For epochs, the maximum value was set to 50, and the optimal loss value was adopted.

MOTION SEGMENTATION RESULTS

Metrics

In this study, we used accuracy rate, edit distance, and segmental F1 score with overlapping threshold rate k (F1@k) as evaluation metrics for the segmentation results. The accuracy rate is the percentage of predicted labels that match with the correct labels for all classification results. Edit distance is the minimum number of operations on the predicted label sequence that match the correct label sequence by insertion, deletion, and replacement, divided by the length of the label sequence. In F1@k, intersection over union (IoU) is calculated for each operation interval of the predicted label sequence, and rate $\geq$k is considered as true positive (TP) and other cases as False Positive (FP). The absence of matching predictive labels for each action interval in the sequence of correct labels is counted as a False Negative (FN). The F1 score is calculated from the precision and recall scores and is expressed by the following equations.

$$Precision = \frac{\mathrm{TP}}{TP + FP} \tag{1}$$

$$Recall = \frac{\mathrm{TP}}{TP + FN} \tag{2}$$

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Table 1. Evaluation results divided by task and by whether the data were the same as those of participants in the training data.

Task	Learned / Unlearned person	Acc.	Edit	F1@0.1	F1@0.25	F1@0.5
1	Unlearned	0.973	0.975	0.986	0.985	0.975
2	Learned	0.866	0.886	0.930	0.918	0.881
3	Learned	0.823	0.905	0.917	0.911	0.900
2	Unlearned	0.873	0.868	0.922	0.922	0.868
3	Unlearned	0.795	0.908	0.908	0.908	0.886

Table 2. Average rated confusion matrix (Task 1).

Task 1		Prediction					
		Left	Right	Forward	Back	Stop	Empty
Ground truth	Left	**0.981**	0.000	0.001	0.001	0.000	0.017
	Right	0.000	**0.813**	0.006	0.017	0.000	0.164
	Forward	0.001	0.000	**0.990**	0.000	0.000	0.009
	Back	0.001	0.000	0.001	**0.968**	0.000	0.030
	Stop	0.000	0.000	0.000	0.000	**0.994**	0.006
	Empty	0.017	0.006	0.009	0.012	0.017	**0.938**

Table 3. Average rated confusion matrix (Task 2).

Task 2		Prediction					
		Left	Right	Forward	Back	Stop	Empty
Ground truth	Left	**0.962**	0.000	0.000	0.003	0.000	0.035
	Right	0.217	**0.581**	0.009	0.046	0.000	0.147
	Forward	0.006	0.001	**0.977**	0.000	0.000	0.015
	Back	0.000	0.001	0.000	**0.980**	0.000	0.019
	Stop	0.000	0.000	0.000	0.000	**0.980**	0.020
	Empty	0.026	0.010	0.016	0.016	0.038	**0.894**

Table 4. Average rated confusion matrix (Task 3).

Task 3		Prediction					
		Left	Right	Forward	Back	Stop	Empty
Ground truth	Left	**0.961**	0.000	0.000	0.006	0.000	0.033
	Right	-	-	-	-	-	-
	Forward	0.004	0.001	**0.717**	0.261	0.000	0.017
	Back	0.000	0.000	0.000	**0.946**	0.000	0.053
	Stop	0.000	0.000	0.000	0.000	**0.982**	0.018
	Empty	0.023	0.007	0.013	0.029	0.041	**0.887**

While F1@k penalizes excessive segmentation, it does not penalize slight deviations between the predicted label sequence and correct sequence, which makes it useful for evaluating segmentation tasks (Lea et al. 2017).

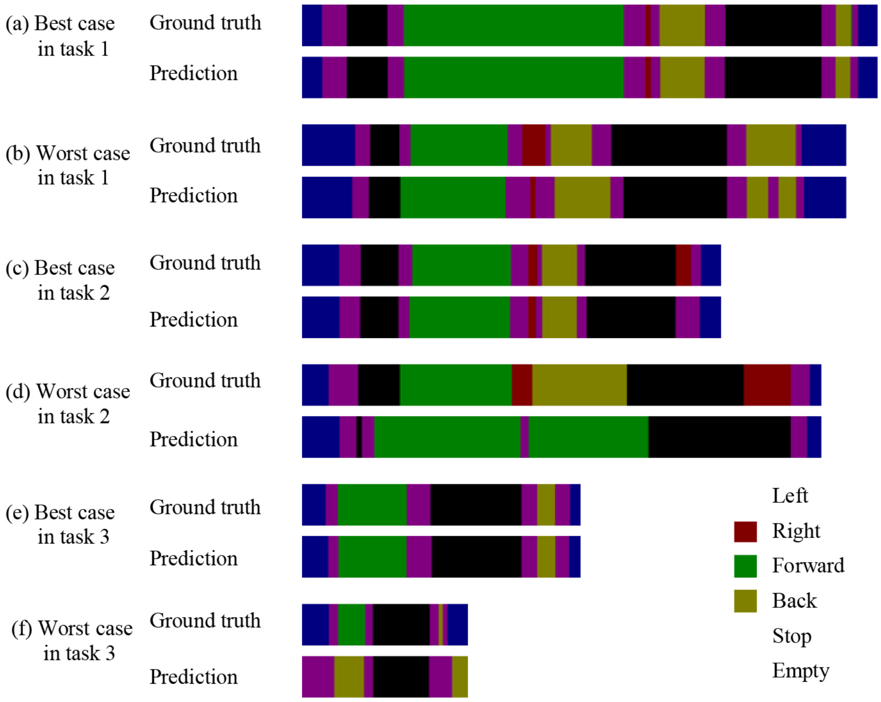

Figure 6: Best and worst accuracy cases for each task.

Motion Segmentation Results

To evaluate the action segmentation results, each evaluation index was calculated based on the results for the test data, and the mean values in the 10-fold cross-validation were calculated. Values of 0.1, 0.25, and 0.5 were used as thresholds for F1@k. The results show a accuracy rate of 0.901, edit distance of 0.923, F1@0.1 of 0.951, F1@0.25 of 0.947, and F1@0.5 of 0.917.

To evaluate the predictive ability of the test data for the people and procedures that were not included in the training data, we divided the test data not only by task but also by whether the data were the same as those of the participants in the training data. Table 1 lists the mean values for the 10-fold cross-validation. Here, the evaluation values were calculated for each datum.

Moreover, Tables 2, 3, and 4 list the results of each cross-validation averaged by scaling each element of the confusion matrix for each task by the total number of correct labels in each class. In all cases, the correct prediction was the largest for each class. However, the percentage of Right misrecognized as Left for Task 2 was 0.217, and the percentage of Forward misrecognized as Back for Task 3 was 0.261, which are relatively high values.

Figure 6 shows the highest and lowest percentages of correct answers for each task.

DISCUSSION

Despite the small number of training data (90) and the existence of unlearned tasks with different procedures, the results of this study showed that not only the percentage of correct answers but also edit distance and each F1@k were equal to or higher than accuracy rate, suggesting the possibility of motion segmentation applications in work training and education.

Table 1 shows that the prediction results for Task 1 for the participant trained on the model showed high values exceeding 0.97 for all indices. This suggests that recognition is possible even for participants who are not present in the training data. This may be due to the fact that the feature extractor is able to adequately capture the characteristics of movements rather than persons through pre-training. In contrast, Tasks 2 and 3 for learned persons showed lower overall predictions than Task 1 for unlearned persons. Table 3 shows that in Task 2, the percentage of Right misrecognized as Left was 0.217, which is larger than that of the other classes. First, the two classes had the same handles to operate, and the difference was only the direction. Next, from in Figure 6 (d), Task 2 includes an action that is not present in Task 1 and becomes Right immediately after Left, confirming that Right was misrecognized as Left there. These two factors may have contributed to the misrecognition of Right as Left in Task 2. The above suggests that misrecognition of unlearned procedures is an issue, even in a pre-trained model.

Table 4 shows that in Task 3, the percentage of Forward misidentification as Back was 0.261, which was larger than that of the other classes, while the reverse misrecognition did not occur at all. This may be due to the fact that Task 1, which was used to train the model, involved facing, whereas Task 3, which was used to train the model, only involved moving the carriage. In this regard, it is necessary to subdivide the labeling and separate the movements that only move the table from those that perform cutting.

CONCLUSION

In this study, we applied an action segmentation model to three types of cutting operations using a lathe in VR space. To develop a method to automatically recognize tasks and actions for more efficient education and training, the results were evaluated on data of persons and actions that were not used for training. Correctness rate, edit distance, and segmental F1 score were used as evaluation indices, and the overall validation showed that all indices were above 0.9, suggesting that the action segmentation model is useful for recognizing tasks and actions even for small training data. The values of each index remained above 0.9 even for human actions that were not used in the training, suggesting generalizability to different people. However, some issues were identified, such as performance degradation for actions of different procedures.

REFERENCES

Carreira, J. Zisserman, A. (2017). “Quo vadis, action recognition? a new model and the kinetics dataset”, proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308, Honolulu, Hawaii.

Farha, Y.A. Gall, J. (2019). “MS-TCN: Multi-stage temporal convolutional network for action segmentation”, proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3575–3584, Long Beach, California.

Ishikawa, Y. Kasai, S. Aoki, Y. Kataoka, H. (2021). “Alleviating over-segmentation errors by detecting action boundaries”, proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2322–2331, Virtual.

Lea, C. Flynn, M.D. Vidal, R. Reiter, A. Hager, G.D. (2017). “Temporal convolutional networks for action segmentation and detection”, proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156–165, Honolulu, Hawaii.

Ministry of Economy, Trade and Industry (METI). (2021) The White Paper on Monodzukuri 2021. METI Website: https://www.meti.go.jp/report/whitepaper/mono/2021/pdf/all.pdf (in Japanese)