Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-023-02152-x

Fine-grained activity classification in assembly based on multi-visual modalities

Haodong Chen[1] · Niloofar Zendehdel[1] · Ming C. Leu[1] · Zhaozheng Yin[2]

Received: 17 November 2022 / Accepted: 19 May 2023

Abstract
Assembly activity recognition and prediction help to improve productivity, quality control, and safety measures in smart factories. This study aims to sense, recognize, and predict a worker's continuous fine-grained assembly activities in a manufacturing platform. We propose a two-stage network for workers' fine-grained activity classification by leveraging scene-level and temporal-level activity features. The first stage is a feature awareness block that extracts scene-level features from multi-visual modalities, including red–green–blue (RGB) and hand skeleton frames. We use the transfer learning method in the first stage and compare three different pre-trained feature extraction models. Then, we transmit the feature information from the first stage to the second stage to learn the temporal-level features of activities. The second stage consists of the Recurrent Neural Network (RNN) layers and a final classifier. We compare the performance of two different RNNs in the second stage, including the Long Short-Term Memory (LSTM) and the Gated Recurrent Unit (GRU). The partial video observation method is used in the prediction of fine-grained activities. In the experiments using the trimmed activity videos, our model achieves an accuracy of > 99% on our dataset and > 98% on the public dataset UCF 101, outperforming the state-of-the-art models. The prediction model achieves an accuracy of > 97% in predicting activity labels using 50% of the onset activity video information. In the experiments using an untrimmed video with continuous assembly activities, we combine our recognition and prediction models and achieve an accuracy of > 91% in real time, surpassing the state-of-the-art models for the recognition of continuous assembly activities.

Keywords Fine-grained activity · Activity classification · Assembly · Multi-visual modality

Introduction

Activity recognition on the industry floor automatically detects and classifies different activities in manufacturing environments. The goal of activity recognition is to understand the nature of the work environment, which allows for a better understanding of how people perform their jobs and what they are operating at any given time (Ahn et al., 2023; Chen et al., 2021; Rude et al., 2018; Xiao et al., 2022). Our previous work analyzed the recognition of coarse-grained gestures (Chen et al. 2022, Chen et al. 2020a, Al-Amin et al., 2021) and worker assembly operation steps (Tao et al., 2020). While activity recognition is a current focus of research, the industry's challenging problem of fine-grained activity recognition is largely overlooked. In industry assembly, fine-grained activity recognition needs to identify similar activities with low inter-class variability and determine the exact type of operations/activities that a worker is doing, rather than recognize coarse activities or assembly steps with high inter-class variability. For example, a coarse-grained machine assembly activity in the industry is annotated as a sequence of elementary sub-actions derived from five fine-grained activity sets: "take the needed parts," "take the needed tools," "put on the assembly platform," "assemble parts," and "check the connection between parts/assembly orientation". The recognition of fine-grained activities in the industry can be used to make more informed decisions about how best to allocate resources, improve quality, and reduce costs. These

✉ Haodong Chen
h.chen@mst.edu

[1] Department of Mechanical and Aerospace Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA

[2] Department of Biomedical Informatics & Department of Computer Science, Stony Brook University, Stony Brook, NY 117943, USA

Published online: 08 June 2023

allow companies to assign workers to specific tasks and ensure they are working on the right things at the right time (Sherafat et al., 2020; Zheng et al., 2021). To realize robust and accurate recognition of fine-grained activity in assembly, we propose a two-stage approach for workers' fine-grained activity recognition by leveraging feature information from the scene-level and temporal-level. We then use the proposed model to recognize and predict fine-grained activities in assembly.

Related work

Traditional coarse-grained activities are concerned with scene-level information, usually involving discrete activities with highly distinct inter-class features. They do not include detailed features on continuous activities in applications (Hu et al., 2020). These highly distinct inter-class features are presented in many popular benchmark activity datasets, such as KTH (Schuldt et al., 2004), UT-interaction (Ryoo & Aggarwal, 2009), and UTKinect-Action3D (Xia et al., 2012). These datasets mainly contain simple daily activities, mostly full-body, such as jogging and hand-waving. Though some fine-grained activity datasets, such as UCF101 (Soomro et al., 2012a), MPII Cooking (Rohrbach et al., 2012) and CAP (Byrne et al., 2023), focus on fine-grained activities sharing similar inter-class features, the public fine-grained activity datasets comprising activities in manufacturing assembly are missing. In this paper, we analyze the complex assembly process of a carving machine and propose a dataset of 15 commonly existing fine-grained activities, which helps fill the above knowledge gap of the fine-grained activity dataset in assembly.

In recognition of fine-grained activities with similar features, some research work has been done. Singh et al. (Singh et al., 2016) presented a multi-stream bi-directional network for fine-grained activity detection. The authors used a bounding box around the subject to avoid most background noise and achieved 80.31% accuracy on public dataset recognition. Pan et al. (Pan et al., 2020) proposed a heterogeneous cyber-physical system using vibration and electrical sensors to monitor fine-grained activities of daily living, which achieved an average of 90% recognition accuracy. Furthermore, some research work has been devoted to fine-grained activity recognition using contextual information between humans and objects (Kapidis et al., 2019; Li et al., 2020; Marszalek et al., 2009; Yao et al., 2011). An assembly fine-grained activity recognizer should distinguish an individual activity from others using the temporal-level features because the background information is usually similar, and the temporal features reflect the dynamic activity process. However, most activity recognition models' performances are often dominated by spatial information, making it challenging to obtain a decent model for fine-grained activity recognition. To address the above issues, we propose a two-stage approach to classify fine-grained activities by combining scene-level and temporal-level features, and shifting the model's attention to temporal features using Recurrent Neural Network (RNN) methods.

Generally, a deep-learning-based recognition task needs a large amount of data to train a deep-learning model, which is time-consuming. For a small dataset, the transfer learning has been demonstrated to be an effective and efficient approach to transfer learning abilities from pre-trained source models to target models (Chan et al., 2023). Kumar et al. (Kumar & Gupta, 2023) investigated the efficacy of transfer learning approaches for predicting various eye diseases and proposed multiple transfer learning models based on limited image data of eye diseases for disease prediction. Fu et al. (Fu et al., 2021) used transfer learning to recognize human activity based on inertial measurement unit (IMU) sensors. That study directly transferred unlabeled data to the model based on the unsupervised method. Transfer learning is also widely used in emotion recognition (Akhand et al., 2021), speech recognition (Cho et al., 2018), medical skin lesion detection (Khan et al., 2021), etc. However, most research applied transfer learning by simply freezing layers before the fully connected layer and adding a customized pooling or fully connected layer. Such an operation may limit the performance of the pre-trained model because transfer learning's performance may vary for different model architectures and use cases. To find the optimal pre-trained model and pooling method for fine-grained recognition in assembly, we compare the performance of different pre-trained models in the transfer learning and evaluate the effect of different pooling methods on the fine-grained activity recognition.

The application and deployment of fine-grained activity recognition necessitate the recognizer to detect the continuously changing activities and provide correct results simultaneously. Mekruksavanich et al. (Mekruksavanich & Jitpattanakul, 2022) introduces a new framework for recognizing sport-related fine-grained activity using multimodal wearable sensors in multiple body positions. The experimental results showed that the proposed recognition model achieved an accuracy of 99.62% on the UCI-DSADS dataset. Zhang et al. shed light on fast action recognition by lifting the reliance on the optical flow and achieved 97.2% accuracy in the experiment on the UCF101 dataset (Zhang et al., 2020a). Kobayashi et al. recognized assembly actions by extracting hand features in two different ways, including cutting out the hand image of a worker and applying an attention module, respectively. The dataset included 11 discrete and low inter-class similarity assembly activities (Kobayashi et al., 2019). Jones developed an assembly action recognition framework with the notion of a kinematic state and defined an action as a difference between two kinematic states to recognize the assembly actions in the constructions of an

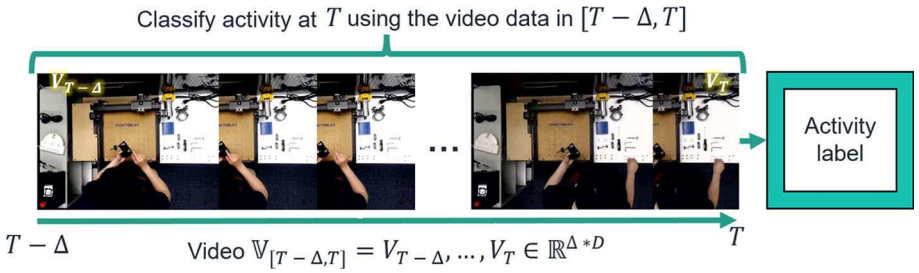


Fig. 1 Problem statement of the fine-grained activities recognition

IKEA chair and toy blocks (Jones et al., 2021). However, in recognition of continuously changing fine-grained activities in an untrimmed video, the inter-class feature variations of the activities are lower than the pre-trimmed video dataset, such as HMDB51 and UCF101, increasing difficulties in the recognition. Additionally, most existing models are not sensitive enough to capture the random emergence of continuous assembly activities of different duration time and unfixed motion speeds. These issues limit the recognition performance of continuously changing activities with an unknown duration time in an untrimmed video. To improve the above issues, we propose a fusion recognition-prediction model based on partial video observation to provide high-accuracy performance in continuous fine-grained activity recognition.

Contribution of this article

To sense and classify the worker's continuous fine-grained activities in manufacturing assembly, in this paper we analyze the fine-grained activities in assembly and propose a network combining the features of Red-Green-and-Blue (RGB) and hand skeleton frames.

The main contributions of this work are as follows:

- A dataset consisting of 15 fine-grained activities in the assembly has been created. This dataset fills in the gaps of manufacturing activity data in existing datasets. This dataset will be shared with the community via our GitHub website.
- A two-stage architecture has been designed in the network to recognize workers' fine-grained activities by leveraging scene-level features in the first stage and temporal-level features in the second stage of the model. The effects of different input lengths, pre-trained feature extractors, RNN models, and fusion mechanisms on fine-grained activity recognition are compared.
- A fine-grained prediction approach using partial video observation to predict upcoming activities is proposed, and a fusion recognition-prediction model is designed.
- Experiments have been conducted to evaluate our model and demonstrate its effectiveness in recognizing and predicting continuous fine-grained assembly activities in real time. Our model is compared with the state-of-the-art models to demonstrate its superiority and generalization capabilities.

The remainder of this paper is organized as follows. Section Problem statement gives the problem statement of this study. Section Dataset collection presents the dataset collection and construction of the fine-grained activities in assembly. Section Two-stage network architecture details our proposed classification model architecture. The experimental setups and results are described and illustrated in Sect. Experiments and results. Finally, Sect. Conclusion presents the conclusion.

Problem statement

As shown in Fig. 1, suppose the obtained fine-grained activity video data is $\mathbb{V}_{[T-\Delta,T]} = [\mathrm{V}_{T-\Delta}, \ldots, \mathrm{V}_T] \in \mathbb{R}^{\Delta * D}$, where $\mathbb{V}_{[T-\Delta,T]}$ represents a video consisting of video frames $[\mathrm{V}_{T-\Delta}, \ldots, \mathrm{V}_T]$ on different time steps $T-\Delta, \ldots, T$. The symbol Δ represents the time interval which goes back to a past time step from the current time step T, while the symbol D is the feature dimension extracted from the camera. Our study aims to classify the continuous fine-grained assembly activities at each time step using the video data in the time interval $[T-\Delta, \ldots, T]$.

Dataset collection

Fine-grained activities in assembly

In the design of a fine-grained activity video dataset of assembly, we analyze the operations in the assembly of a desktop carving machine and extracted 15 fine-grained activities, which are: *assemble parts* (A1), *background* (A2), *check* (A3), *connect cables* (A4), *organize cables* (A5), *place a tool/part* (A6), *push a button* (A7), *put a part on board* (A8), *take parts/tools from the left* (A9), *take parts/tools from the right* (A10), *tighten by hands* (A11), *untie cables* (A12), *use a screwdriver* (A13), *use a small Allen key* (A14), *use a big Allen key and a wrench* (A15). These 15 operations are

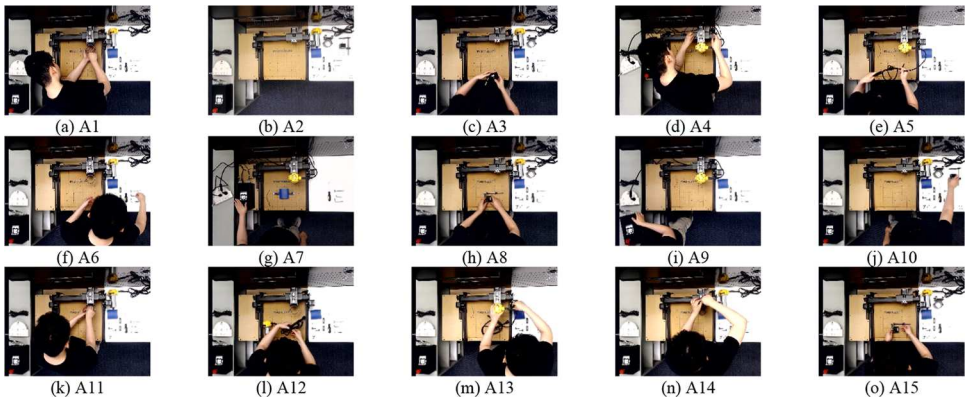


Fig. 2 Fine-grained activities in the assembly of a carving machine

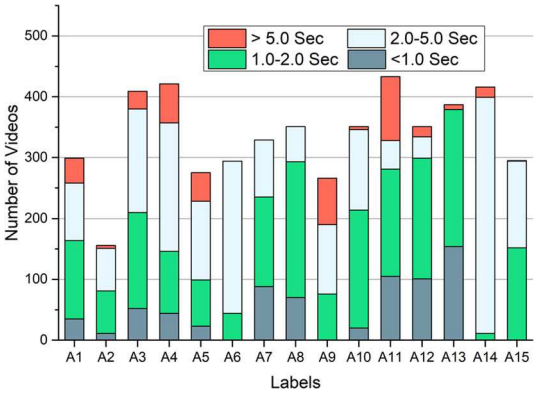


Fig. 3 Quantities of the proposed fine-grained activity dataset in assembly

shown in Fig. 2. Particularly, since the activities of taking different tools or parts have the same activity characteristics, we grouped them into two categories based on the location of the parts: *take parts/tools from the left* (A9) and *take parts/tools from the right* (A10). The classes of the *assemble parts (*A1*)*, *place a tool/part* (A6), and *put a part on board* (A8) were designed based on similar rules. One Logitech C920 camera is used in data collection, with a frame resolution of 1920 × 1080 pixels and a frame rate of 30 fps. Three subjects are involved in dataset construction. The dataset will be shared with the community via our GitHub website. The quantity of the video dataset is shown in Fig. 3. The total time and average clip time of the video dataset are shown in Fig. 4.

Hand skeleton frame extraction

We obtain the RGB frames from the video dataset and then extract the hand skeleton frames from the RGB frames using the MediaPipe approach (Zhang et al., 2020b). The process is shown in Fig. 5, in which the approach checks each frame and detects if hands exist. The landmark model determines the precise location of the 21 hand-knuckle coordinates inside the detected hand regions and draws the hand skeletons. In our work, we draw the hand skeletons on a black background of the same size as the input frame to shift the model's attention to hand movement features. The MediaPipe approach achieves an average accuracy of ≈ 96% in palm detection. In cases the hand skeletons are not correctly detected in an activity, as shown in Fig. 6, our model in Sect. Two-stage network architecture can use the temporal-level features of the continuous frames to eliminate the effect of the hand skeletons missing in a single frame.

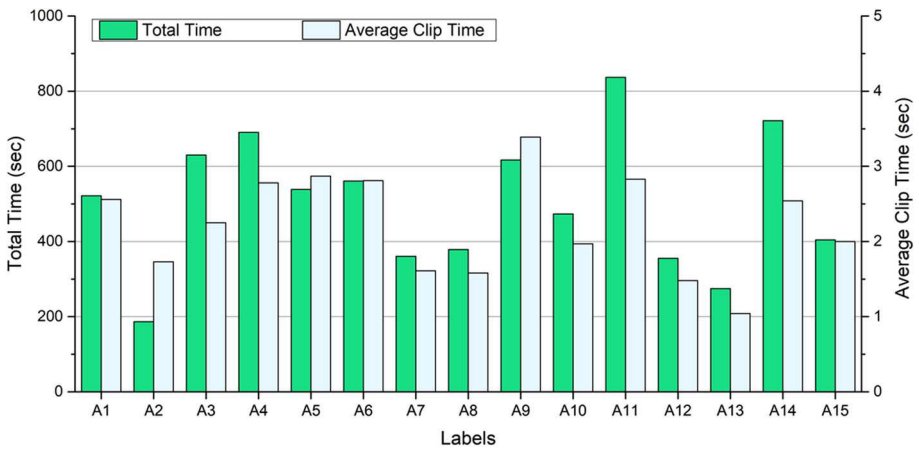


Fig. 4 Duration time of the proposed fine-grained activity dataset in assembly

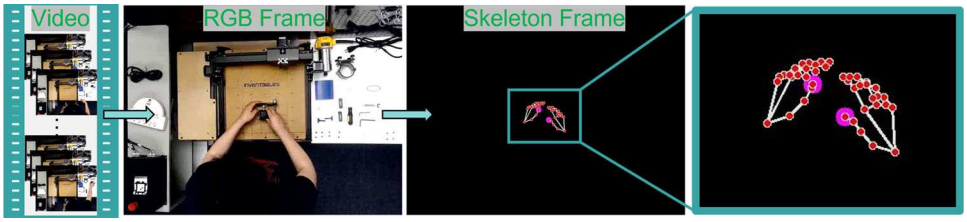


Fig. 5 Hand skeleton frame extraction

Fig. 6 Samples of hand skeleton frame extraction

Two-stage network architecture

Data normalization

Different activities usually have different lengths in a video dataset. The video sequences are normalized into the same length T. Samples longer than the length T are all evenly truncated, and samples shorter than the length T are padded with the missing frames. The padded frames are marked as false in the feature extraction and training since they are used to meet the length requirements (https://www.tensorflow.org/guide/keras/understanding_masking_and_padding). To analyze the impact of different input sequence lengths on fine-grained activity recognition and prediction, we process the activity video data into five fixed input lengths, including 10, 20, 40, 60, and 80 frames. The experimental results are shown in Sect. Evaluation of different pre-trained models and RNNs.

Network architecture

Our two-stage network is shown in Fig. 7. There are five components, including the input video, multi-visual frames, first stage, second stage, and output of the classification.

Input video: The frame rate of the input video is 30 fps.

Multi-Visual frames: The RGB and skeleton frames are first extracted and normalized into the same dimension, including the dimension of the individual frame and the length of the frame sequence.

First stage: We use transfer learning to extract the scene-level features of the normalized RGB and skeleton frames. The source dataset of the pre-trained models contains many annotated data, with which a deep learning model is trained. After that, a portion of the pre-trained model and the trained weights are frozen and transferred to our target domain, i.e., the fine-grained activities. A new classifier is designed to adapt the source model to the target domain. To create an end-to-end neural network, we removed the last fully connected layer of the pre-trained model in the first stage and directly connected the output to the second stage.

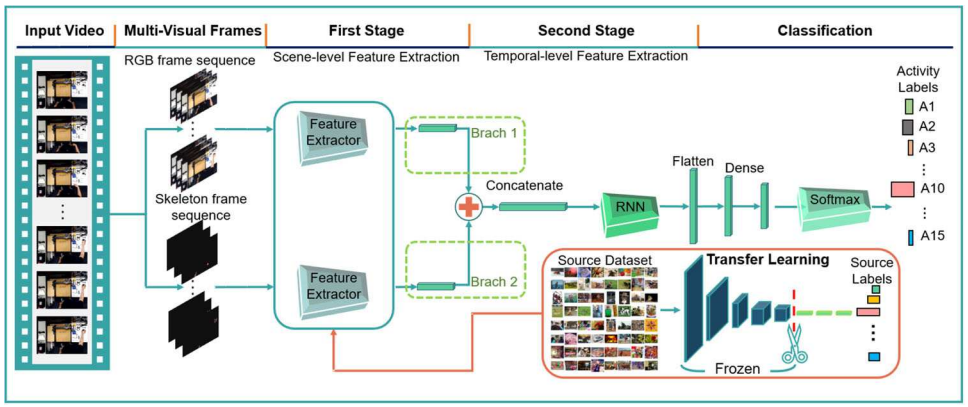
Second stage: After the scene-level feature extraction, the obtained features are input to an RNN layer, which extracts the temporal-level features of the input sequence and combines them with the scene-level features of the RGB and skeleton frames. We propose two fusion strategies to fuse the two feature modalities before (Fig. 7a) and after (Fig. 7b) the RNN layer. The model fused before the RNN layer concatenates the unimodal features in the two branches in Fig. 7a into a single representation, and an upcoming individual RNN layer merges the temporal-level features of the RGB and skeleton frame sequences. The model fused after the RNNs in Fig. 7b learns the RGB and skeleton frame sequence's temporal-level features separately and then concatenates the two unimodal branches of the two RNN layers in Fig. 7b.

Output: The classification of the fine-grained activities follows the first and second stages. After a Softmax layer, the activity label is output. The Softmax layer is shown in Eq. (1):

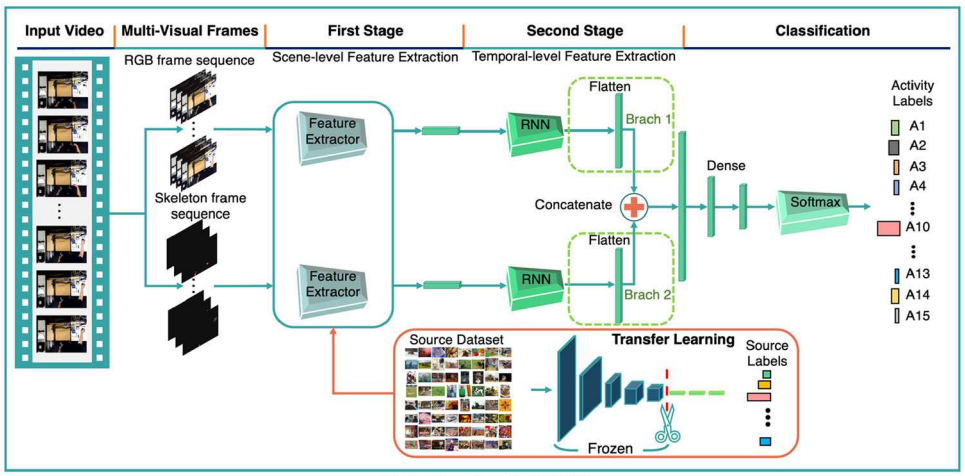
$$P(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{15} e^{x_k}} \text{ for } i = 1, \ldots, C \tag{1}$$

where C represents the number of the classes, and $C = 15$ in the assembly fine-grained activity recognition. The above Softmax layer limits the output value to the interval [0, 1] and makes the sum of components to be 1 so that the output of the Softmax layer can be interpreted as recognition probabilities (Chen et al., 2020b).

Regarding the length of the input video, we analyze the impact of different input lengths on fine-grained activity recognition. We process the activity video data into five fixed input lengths, including 10, 20, 40, 60, and 80 frames. The experimental results of the five input lengths are given in Sect. Evaluation of different pre-trained models and RNNs. In the first stage, we freeze the layers of the pre-trained model before the fully connected layers at the decision level and



(a) Fusion before the RNN layer



(b) Fusion after the RNN layers

Fig. 7 The architecture of the proposed models

(a) Original (b) Maximum pooling (c) Average pooling

Fig. 8 Samples of maximum pooling and average pooling

transfer the source domain to the target domain, i.e., the fine-grained activities. A pooling layer is added after the frozen layers to reduce the size of feature maps by summarizing the presence of features in each patch of the feature map. There are two popular pooling methods, i.e., maximum pooling and average pooling, as shown in Fig. 8b and c. The maximum pooling method calculates the maximum value in each frame feature map patch and replaces each element in the patch with the maximum value. The average pooling method calculates the average value for each frame feature map patch and replaces each element in the patch with the average value. We compare the performance of the two pooling methods in Fig. 8 using 10% of the dataset with the three pre-trained models ResNet50 (He et al., 2016), VGG-16 (Simonyan & Zisserman, 2014a), and InceptionV3 (Szegedy et al., 2016). The experimental results will be presented in Sect. Evaluation of different pre-trained models and RNNs. We conducted an additional experiment in Sect. Evaluation of different pre–trained models and RNNs 1 to evaluate the performance of using different pre-trained models as feature extractors on RGB and skeleton frame sequences separately.

In transfer learning, we addressed the "dated weight" problem caused by Batch Normalization (BN) layers in pre-trained models, which are trained on the ImageNet dataset with significantly higher inter-class variation than our dataset, resulting in degraded performance. To address

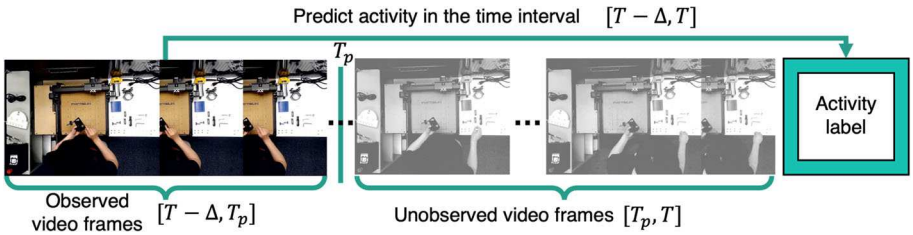


Fig. 9 Fine-grained activity prediction using partial video observation

this issue, we applied the Batch Renormalization technique (Tian et al., 2020) in the first stage of our model, which modifies the scaling and shifting parameters of the BN layers to adapt to changes in the input data distribution, making it less dependent on the mini-batch size and better suited for our fine-grained activity dataset.

The RNN is used to process and recognize sequential activity data. It builds deep neural networks with recurrent architectures and effectively solves recognition problems involving sequential data. The input data is first transformed into machine-readable vectors. Then the RNN processes the vector sequence one by one. While processing, it saves a previous input as a hidden state, passes it to the next sequence step, and merges the current and previous inputs into a new hidden state. The hidden state acts as the memory of the neural network. It holds information on previous data seen by the network. Two popular RNN models are used in our task; they are the Long Short-Term Memory (LSTM) (Yu et al., 2019) and the Gated Recurrent Unit (GRU) (Cho et al., 2014).

Prediction of fine-grained activities using partial video observation

We design action prediction as shown in Fig. 9, in which the activity frames are partially observed and used to train a prediction model. The prediction model shares the same architecture as the model in Sect. Network architecture, but uses only the beginning frames to partially learn the activity features, i.e., using the activity frames in the time interval $[T-\Delta, T_p]$ to identify the activity in the time interval $[T-\Delta, T]$, where $T_p < T$. Different partial ratios are tested, including 50%, 25%, and 12.5% of temporal information for given activity video samples. The experimental results of the three cases mentioned above are presented in Sect. Evaluation of the prediction using trimmed videos.

Experiments and results

The experimental platform is a workstation with an Ubuntu 16.04 system equipped with 64343 M RAM and an NVIDIA GeForce RTX 3090 graphics card. The dataset records ten complete assembly processes of three human subjects, and there are more than 5000 activity samples. We randomly divided the dataset into training, validation, and testing sets using a 6:2:2 split, where 60% of the data was used for training, 20% for validation, and 20% for testing. Table 1 presents the training parameters used in this study, including the learning rate, decay for the update of the learning rate in each iteration, batch size, and the number of epochs.

Table 1 Training parameters used in the proposed model

Learning Rate	Decay	Batch Size	Epoch
1e-5	1e-6	32	200

Two types of videos are used in the experiments: trimmed and untrimmed. Each video contains only one fine-grained activity in the trimmed videos, while the untrimmed video records a completed assembly process, including hundreds of continuous fine-grained activities. The following experiments are carried out: (i) evaluation of the recognition performance under three pre-trained models (ResNet50, VGG-16, and InceptionV3) and two RNN models (the LSTM and GRU), (ii) evaluation of the recognition performance of the proposed model using the trimmed activity videos, (iii) evaluation of the prediction performance of the proposed model using the trimmed activity videos, (iv) evaluation of the recognition performance of the proposed model using the untrimmed activity videos, and, (v) comparison with the recognition results using state-of-the-art models from the literature. Several widely used metrics are used to assess classification performance:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

where the True Positive (TP) refers to a sample x belonging to a class C that is correctly classified as C. True Negative (TN) indicates that a sample x from a 'not C' class is correctly classified as a member of the 'not C' class. The False Positive

(*FP*) is when a sample x from a 'not *C*' class is incorrectly classified as class *C*. The False Negative (*FN*) describes a situation, in which a sample x from class *C* is misclassified as belonging to 'not *C*' classes. The $F_1 Score$ is the harmonic mean of the *Precision* and *Recall*, which ranges in the interval [0, 1].

Specifically, we used a learning rate of 1e-5 and a decay of 1e-6 for the update of the learning rate in each iteration. Additionally, we set the batch size to 32 and trained the model for a total of 200 epochs.

Evaluation of different pre-trained models and RNNs

We compare the performance of the maximum and average pooling methods in Fig. 8 using 10% of our dataset with the different pre-trained models ResNet50, VGG-16, and InceptionV3. The experimental results are given in Table 2, where the average accuracy values of LSTM and GRU are given in different cases. We find that the maximum pooling outperforms the average pooling by 2.92% higher in accuracy. Thus, the maximum pooling is used in our transfer learning process.

We conducted experiments to evaluate the impact of using different feature extractors on RGB and skeleton frame sequences. Specifically, we compared the performance of three different feature extractors (VGG16, ResNet50, and InceptionV3) for RGB frames and skeleton frames using 10% of our dataset. The results, presented in Fig. 10 where the rows and columns represented the feature extractors of RGB and skeleton frames, respectively, indicate that using the same feature extractor for both RGB and skeleton frames resulted in a 3.67% improvement in accuracy compared to using different feature extractors. The VGG-16 pre-trained model yields better results than ResNet50 and InceptionV3. Therefore, we have selected the VGG-16 as the feature extractor of our model.

Tables 3 and 4 show the recognition accuracy of the three pre-trained models, i.e., ResNet50, VGG-16, and InceptionV3, in four cases, i.e., fusion before and after LSTM, and fusion before and after GRU. The input lengths Δ include 10, 20, 40, 60, and 80 frames, as discussed in Sect. Data normalization. We find that: (i) The recognition accuracy is > 91% in all cases. In the case of the 20-frame input length, VGG-16 pre-trained model, and late-fusion LSTM mechanism, the highest accuracy of 99.63% is achieved. (ii) Compared with the ResNet50 and InceptionV3, the VGG-16 gives a more desirable performance with a recognition accuracy of > 98% in all cases, on average 4% and 3% higher than the results of the ResNet50 and InceptionV3, respectively. (iii) The fusion-after-RNN mechanism performs better than the fusion-before-RNN mechanism, with an average of 2% higher accuracy. (iv) The LSTM performs better than the GRU with 2% higher accuracy. Based on the above results, our optimal model for recognizing fine-grained activities uses the fusion-after-LSTM mechanism shown in Fig. 7b.

The input length Δ in the time interval $[T - \Delta, T]$, in Sect. Problem statement is determined using a five-fold validation experiment. The training data set is divided into five folds. Each fold is treated as a pseudo-test set in turn, and the other four folds are pseudo-train sets. We calculate the average recognition accuracy of fifteen fine-grained activities for each input length Δ with the VGG-16 feature extractor and

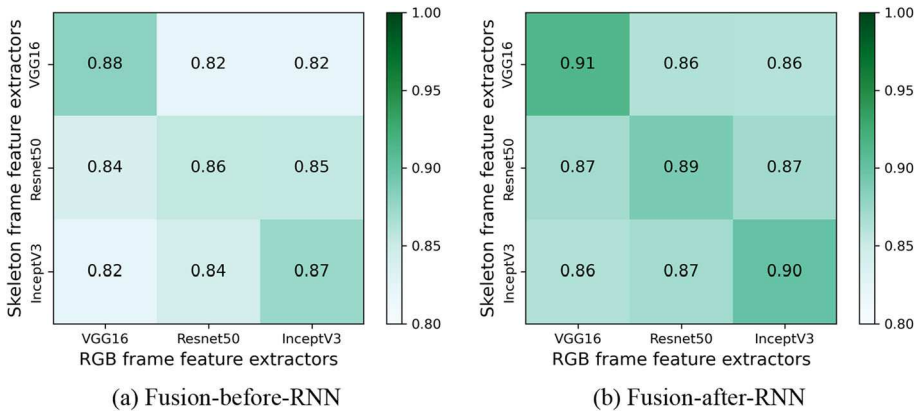


Fig. 10 Performance of different feature extractors on RGB and skeleton frames

Table 2 Performance (%) of the two pooling methods using 10% of the dataset

Pooling Type	Fusion Before RNN Layer			Fusion After RNN Layer		
	VGG-16	ResNet50	InceptionV3	VGG-16	ResNet50	InceptionV3
Maximum	88.06	88.54	89.43	91.26	88.93	90.93
Average	87.61	86.51	87.31	88.64	85.41	87.15

Table 3 Accuracy (%) of different pre-trained models under different input lengths using LSTM

Input Length (frame)	Fusion-before-LSTM			Fusion-after-LSTM		
	VGG-16	ResNet50	InceptionV3	VGG-16	ResNet50	InceptionV3
10	97.34	95.54	93.89	97.79	94.24	95.01
20	98.43	95.74	93.89	99.63	95.24	95.41
40	98.68	92.87	92.70	99.26	96.06	97.08
60	98.53	94.04	96.01	99.26	96.87	97.48
80	98.52	93.04	94.94	98.98	95.05	97.40

Table 4 Accuracy (%) of different pre-trained models under different input lengths using GRU

Input Length (frame)	Fusion-before-GRU			Fusion-after-GRU		
	VGG-16	ResNet50	InceptionV3	VGG-16	ResNet50	InceptionV3
10	98.45	91.00	91.03	98.11	93.85	92.85
20	98.52	92.30	92.86	98.79	93.85	94.86
40	98.89	91.08	93.04	98.53	94.31	94.86
60	98.15	91.88	94.97	98.43	95.02	96.55
80	98.52	92.98	94.98	98.72	95.17	93.45

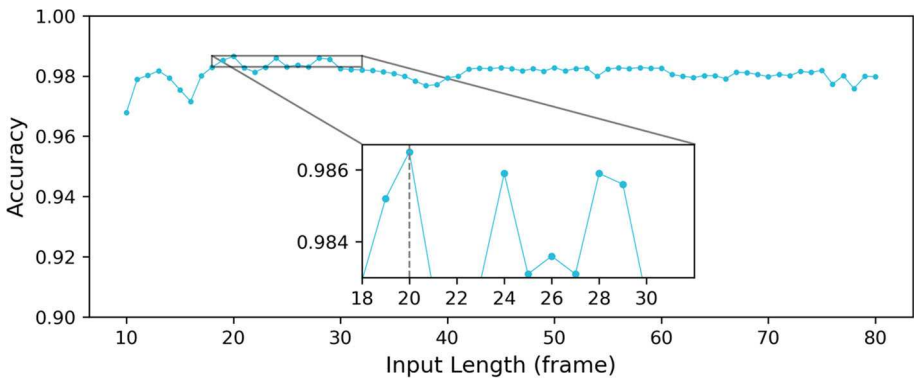


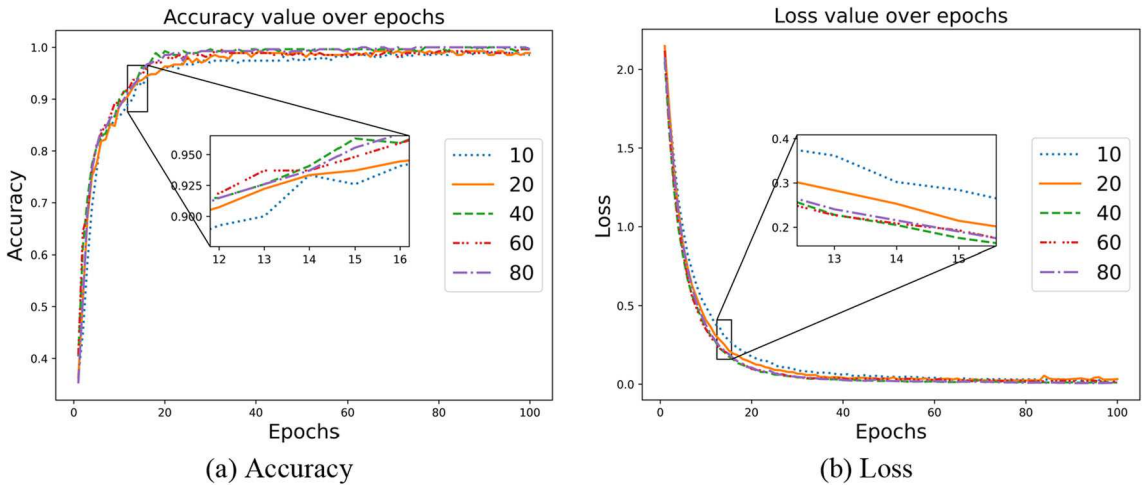
Fig. 11 Performance of different input lengths (Δ) in the five-fold validation experiments using the VGG-16 feature extractor and fusion-after-LSTM mechanism

fusion-after-LSTM mechanism. The results are summarized in Fig. 11, which indicates that the 20-frame input length yields better results than other input frame lengths. Therefore, we have selected the 20-frame as Δ in Fig. 1, i.e., the input length of our model.

To demonstrate the effect of different input lengths on the training process, we plot curves of validation accuracy and loss during the fusion-after-LSTM training process in Fig. 12, in which the horizontal coordinate indicates the number of iterations in the training process and the vertical coordinates represents the accuracy or loss values. We find that: (i) As the number of iterations increases, the accuracy values gradually increase, and the slope gradually decreases for all input length cases in Fig. 12a. The final accuracy values stabilize at around 99% for all input length cases. (ii) As the number of iterations increases, the loss values gradually decrease, and the slope decreases for all input length cases in Fig. 12b. The final loss values for all input length cases stabilize at around 0. (iii) For cases with shorter input lengths, the accuracy and loss converge relatively slowly, e.g., the accuracy and loss of the 10-frame case eventually converge in about 60 iterations. The accuracy and loss convergences are faster for the cases with larger input lengths, e.g., the case with an 80-frame input length converges after about 40 iterations. (iv) The accuracy and loss curves are smooth with no dramatic fluctuations. The accuracy and loss converge to around 99% and 0, respectively, which indicates that we have trained a well-fit model on fine-grained activity recognition.

Evaluation of the recognition using trimmed videos

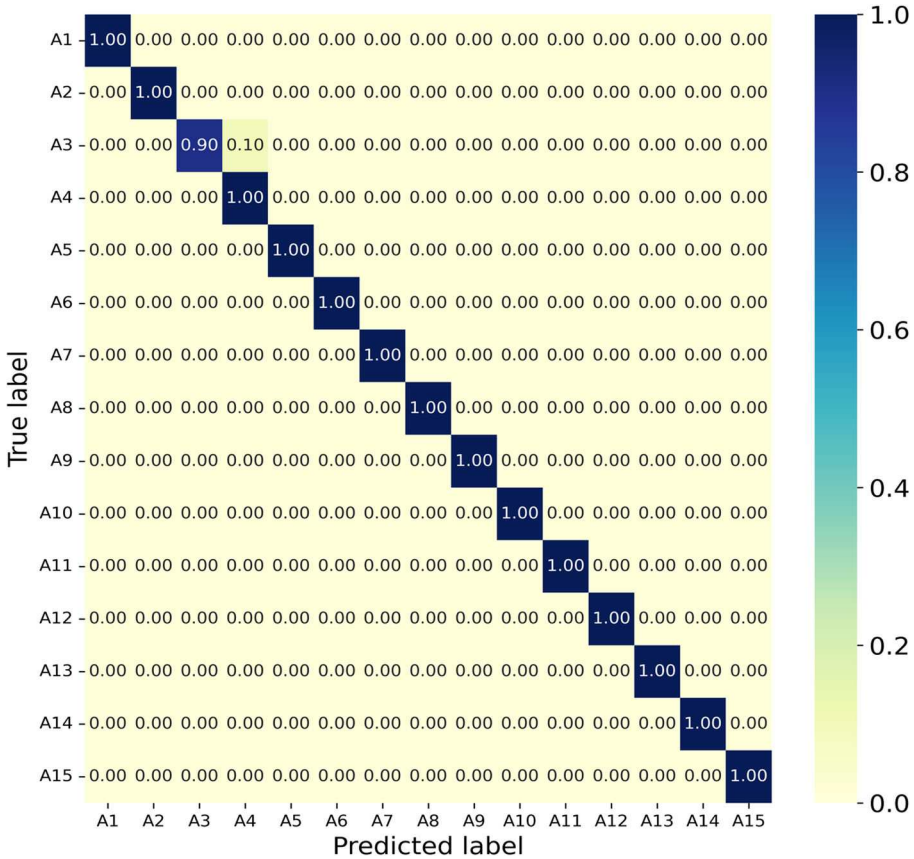
In the trimmed video dataset (Sect. Dataset collection), each video clip includes one activity. The confusion matrix of our optimal model (20-frame input length, VGG-16 pre-trained model, fusion-after-LSTM mechanism) for recognition experiments using the trimmed videos is shown in Fig. 13, where the rows represent the activity ground truth, and the columns represent the predicted labels. In Fig. 13,



(a) Accuracy (b) Loss

Fig. 12 Accuracy and loss curves under different input lengths and the fusion-after-LSTM mechanism

Fig. 13 Confusion matrix under the 20-frame input length



most classification results are concentrated along the diagonal with the accuracy of 100%, showing the high performance of our model, except that 10% of the activity A3 (*check*) is recognized as activity A4 (*connect cables*). The *Accuracy*, *Precision*, *Recall*, and $F_1 Score$ of the model for the 15 activities in Table 5 show the same results, in which all accuracy values are > 99%.

We analyze the failure cases and find that i) more than 95% of misclassified activities A3 are temporally connected to A4 in the original assembly data. These connected actions are trimmed into two categories of video samples with different labels, i.e., A3 and A4. As shown in Fig. 14 A3 (*check*) and A4 (*connect cables)* are two consecutive activities, and the worker naturally checks (A3) the alignment of the cable connections when connecting cables (A4). ii)

Table 5 Performance (%) of the fine-grained activity recognition using trimmed videos (20-frame input length, fusion-after-LSTM mechanism, and VGG-16 pre-trained model)

Activity Class	*Accuracy*	*Precision*	*Recall*	F_1*Score*
A1	100.00	100.00	100.00	100.00
A2	100.00	100.00	100.00	100.00
A3	99.33	100.00	90.00	94.74
A4	99.33	90.91	100.00	95.24
A5	100.00	100.00	100.00	100.00
A6	100.00	100.00	100.00	100.00
A7	100.00	100.00	100.00	100.00
A8	100.00	100.00	100.00	100.00
A9	100.00	100.00	100.00	100.00
A10	100.00	100.00	100.00	100.00
A11	100.00	100.00	100.00	100.00
A12	100.00	100.00	100.00	100.00
A13	100.00	100.00	100.00	100.00
A14	100.00	100.00	100.00	100.00
A15	100.00	100.00	100.00	100.00

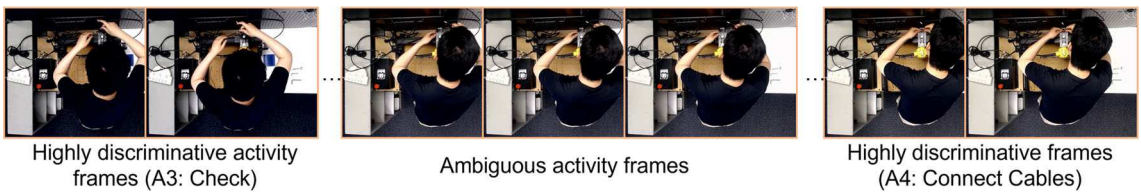


Fig. 14 Highly discriminative and ambiguous activity frames

The fact that activities A3 and A4 share the same operational elements (cables) and hand activity (holding cables) introduces ambiguous frames, which can be defined as both the end of the activity A3 (*check*) and the beginning of the activity A4 (*connect cables*). Over 80% of failure cases of A3 (*check*) do not have ambiguous frames because the corresponding ambiguous frames are trimmed into the subsequent A4 (connect cables), which indicates that most of the correctly recognized activity A3 samples have ambiguous frames. To improve the recognition of activities A3 and A4, we reallocate all the ambiguous activity frames between these two activities to activity A3 (*check*) and increase the recognition *Precision*, *Recall*, and F_1*Score* values of activities A3 and A4 to > 98%.

Evaluation of the prediction using trimmed videos

The performance of the activity prediction using the partial video observation method is shown in Table 6, in which T represents the activity length after video data normalization. We find that: (i) The average accuracy for the cases of 20–80 frame input lengths is 81.96%, 90.48%, and 96.95% when the observation ratio is 12.5%, 25%, and 50%, respectively. The accuracy for the cases with input lengths of 20 and 60 frames is higher than in the other cases at a 50% observation ratio, with the same accuracy of 97.82%, indicating similar prediction performance. (ii) The average recognition accuracy for the cases of 10-frame input length is lower than the cases of other input lengths for different local video observation ratios. It indicates that the input length of 10 frames cannot obtain sufficient activity information. The performance in Table 6 shows that the prediction model using partial video observation is sensitive and robust in predicting fine-grained activity labels.

Table 6 Accuracy (%) of the prediction using trimmed videos

Input Length (frame)	Video Observation Ratios		
	12.50% T	25.00% T	50.00% T
10	67.97	79.90	84.81
20	82.98	90.90	97.82
40	78.52	90.14	96.74
60	83.87	88.97	97.82
80	82.45	91.90	95.24

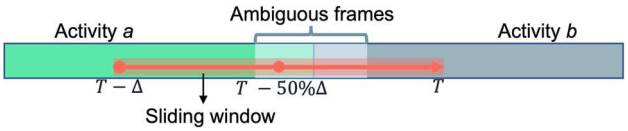


Fig. 15 Continuous fine-grained activity recognition and prediction using an untrimmed video

Evaluation of the recognition and prediction using untrimmed videos

In recognition of continuous fine-grained activities in an untrimmed video, a sliding window with the Δ time length slides between continuous activities. The untrimmed input video consists of 20,640 frames ($\approx$ 11 min) and more than 120 continuous activities in Sect. Fine-grained activities in assembly. We compare three cases shown in Fig. 15: (i) Only using the recognition model with the input frames within the time interval $[T - \Delta, T]$ in the sliding window to continuously recognize the activity at the current time T. (ii) Only using the 50%T prediction model with the input frames within the time interval $[T - 50\%\Delta, T]$ in the sliding window to identify the activity at the current time T. (iii) A model that fuses both the recognition and 50%T prediction models, which combines the recognition result using the activity frames within the time interval $[T - \Delta, T]$ and the prediction result using the activity frames within the time interval $[T - 50\%\Delta, T]$ in the sliding window to identify the activity at the current time T. The fusion model outputs the result with higher confidence as the final recognition result.

To analyze the impact of different recognition frequencies on the continuous recognition of activities with different lengths and speeds, we compare the performance of different recognition frequencies $F = 1$, 3, 6, 10, and 30 Hz at a fixed frame rate, i.e., 30 fps. When using a higher recognition frequency, most activity frames are shared in adjacent recognitions, e.g., a recognition frequency $F = 30$ Hz with a 60-frame input length indicates the recognition is conducted 30 times in 1 s to recognize activities in the past 60 frames, which means that every two adjacent recognition operations share 59 frames. When using a lower recognition frequency, more past frames are discarded in each new recognition, e.g., a recognition frequency $F = 1$ Hz with a 60-frame input length indicates that only one recognition is performed in 1 s and that only 30 frames are shared between every two adjacent recognitions. An optimal recognition frequency with the corresponding input length balances the contribution of past and new input frames to the recognition performance, allowing the model to be more sensitive and robust in recognizing continuous fine-grained activities in real time.

The experimental results of continuous fine-grained activity in an untrimmed video with five input lengths (10, 20, 40, 60, and 80 frames) at five different recognition frequencies (1, 3, 6, 10, and 30 Hz) are given in Table 7, in which A_d and A_a represent the highly discriminative activity frames and ambiguous activity frames, as discussed in Sect. Evaluation of the recognition using trimmed videos. We find that: (i) the cases of 20-frame input length have higher accuracy than the other cases. The highest accuracy results are found for the case of 20-frame input length and 6 Hz recognition frequency, which are 94.35% and 83.85% for the highly discriminative activity frames A_d, and ambiguous activity framesA_a, respectively. This result is consistent with the recognition results using the trimmed videos in Sect. Evaluation of the recognition using trimmed videos, i.e., the case of 20-frame input length achieves the highest recognition accuracy. (ii) The fusion model combining the recognition model and the 50%T prediction model provides better performance than either the recognition model or the 50%T prediction model. The combination of the recognition and prediction model contributes to improving recognition accuracy, which improved the classification of the A_d and A_a by 12.55% and 11.34% on average, respectively. (iii) Based on the values of the A_d and A_a in the untrimmed assembly video, i.e., 71.20% and 28.80%, we calculate the average accuracy as $94.35\% * 71.20\% + 83.85\% * 28.80\% = 91.33\%$, which indicates that our model performs well in recognition of continuous fine-grained activities in the untrimmed video. iv) Our model conducts a fine-grained activity recognition (with 20-frame input length and 6 Hz) in ~ 0.03 s (using the camera input at 30 fps, or 0.03 s per frame), thus the recognition is performed in real time.

Evaluation of different locations of the input time interval $[T - \Delta, T]$ in continuous activities

As discussed for Fig. 4 in Sect. Fine-grained activities in assembly, the lengths of all activity samples are longer than 20 frames, which indicates that the input frame sequence in the input time interval $[T - \Delta, T]$ covers at most two continuous activities if we choose 20-frame input length based on the results in the last section. To evaluate the differences among the three cases in Table 7 i.e., only recognition, only 50%T prediction, and recognition + 50%T prediction, five cases are considered in the continuous fine-grained activity recognition using an untrimmed video, as shown in Fig. 16. The sliding window containing input frame sequence in the time interval $[T - \Delta, T]$ slides from the activity a to activity b, where the activities a and b represent random adjacent activities. More than 100 experiments are conducted for each case in Fig. 16, and the inputs are video samples with two random adjacent activities. The experimental results are given in Table 8. We find that: (i) Fusing the recognition model and the 50%T prediction model provides better performance than using each individual model and achieves an accuracy of > 97% for all cases, which is consistent with

Table 7 Performance (%) of continuous fine-grained activity recognition using an untrimmed video

Input Length (frame)	Recognition Frequency	Only Recognition		Only 50%T Prediction		Recognition + 50%T Prediction	
		A_d	A_a	A_d	A_a	A_d	A_a
10	1	45.27	23.54	19.85	13.52	46.28	24.57
	3	59.37	35.35	25.54	13.17	64.94	32.48
	6	71.41	41.59	25.81	15.57	75.57	52.85
	10	78.42	40.38	28.21	15.74	82.57	41.58
	30	71.94	35.55	35.12	14.15	76.84	36.85
20	1	65.94	32.59	34.21	24.73	77.71	41.53
	3	77.85	58.95	61.11	38.77	82.16	69.50
	6	91.43	75.51	81.21	40.57	94.35	83.85
	10	89.44	71.99	74.32	40.38	93.58	78.58
	30	88.94	68.59	72.21	35.57	92.57	76.05
40	1	65.14	37.59	54.12	32.80	67.77	45.51
	3	79.07	58.26	67.27	52.71	84.89	61.84
	6	87.93	78.52	87.95	74.11	92.16	81.83
	10	87.11	71.68	83.53	65.52	89.57	80.20
	30	88.24	61.05	85.36	51.40	89.93	63.72
60	1	68.67	36.52	52.82	31.58	79.31	43.57
	3	79.60	49.82	64.04	38.90	83.72	51.91
	6	89.61	58.49	77.96	38.19	92.65	59.84
	10	87.21	49.78	74.65	39.49	89.49	54.85
	30	87.23	36.67	73.96	33.21	89.57	46.81
80	1	68.93	38.10	66.42	33.67	68.58	39.61
	3	78.94	35.59	70.76	38.17	83.58	39.31
	6	84.84	43.39	80.67	36.25	92.23	45.60
	10	79.87	41.02	80.34	36.46	91.17	43.23
	30	77.11	36.94	78.42	35.40	85.34	41.58

the results in Table 7. (ii) In the cases 1 and 5, the sliding window covers one activity, and the experimental results are > 99% for the fusion model, which is similar as using trimmed videos and the results are consistent with the results using the trimmed videos in Sect. Evaluation of the recognition using trimmed videos and Evaluation of the prediction using trimmed videos. (iii) In the cases 2–4 where the sliding window covers two activities, the 50%T prediction model provides higher accuracy because the activity frames in the time interval $[T - 50\%\Delta, T]$ shift the focus of the model to the new emerging activity near the time T. As discussed in Sects. Prediction of fine-grained activities using partial video observation and Evaluation of the prediction using trimmed videos, the 50%T prediction model is sensitive to the starting frames of activities and can correctly predict activities with an accuracy of 97.82%. Overall, the recognition model considers the entire activity in the recognition, which achieves accurate results in single activity recognition, and the 50%T prediction model considers the 50% beginning activity in the prediction, which maintain the robustness in dealing with the transition between adjacent activities. The fusion model combines the benefits of the recognition model and the 50%T prediction model and provides the accurate and robust results in the continuous fine-grained activity recognition using an untrimmed video.

To validate the effectiveness of the proposed model, we designed four groups of comparative experiments. We tested the model with different configurations for continuous fine-grained activity recognition using a recognition frequency of 6 Hz, and the experimental results are shown in Table 9. We found that: (i) Group 1, which uses only scene-level features, achieves an accuracy of 58.38%, and Group 2, which combines temporal features with scene-level features, obtains an accuracy of 80.54%. This indicates that the temporal features significantly contribute to the classification of fine-grained activities. (ii) Compared to Group 2, the model of Group 3

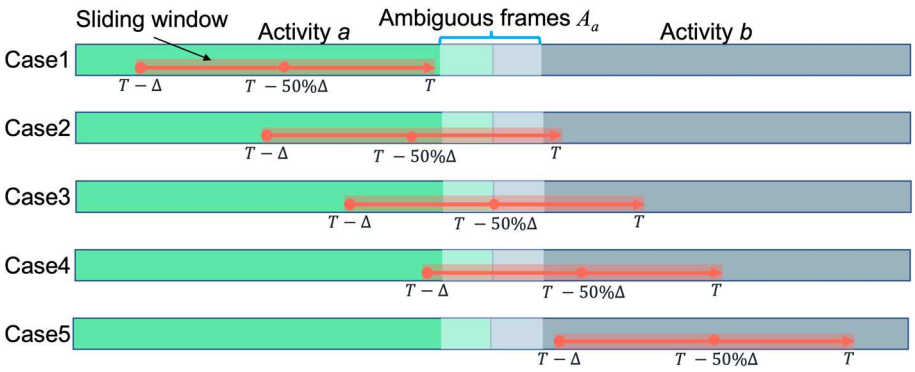


Fig. 16 Five cases in continuous fine-grained activity recognition using an untrimmed video

Table 8 Accuracy (%) of five cases of the sliding window in the time interval $[T-\Delta, T]$

Five Cases in Fig. 16	Recognition		
	Only Recognition (%)	Only 50%T Prediction (%)	Recognition + 50%T Prediction (%)
Case 1	99.05	93.14	99.05
Case 2	88.11	97.65	98.13
Case 3	93.14	98.03	98.06
Case 4	95.15	97.06	99.03
Case 5	99.03	94.12	99.04

Table 9 Results of ablation experiments ($\surd$ means that the strategy is applied, and the $\times$ means that the strategy is not applied)

Group	Scene-level Features	Temporal-level Features	Skeleton Frames	Same Feature Extractors for RGB and Skeleton Frames	50%T Prediction	Accuracy (%)
1	$\surd$	$\times$	$\times$	$\times$	$\times$	58.38
2	$\surd$	$\surd$	$\times$	$\times$	$\times$	80.54
3	$\surd$	$\surd$	$\surd$	$\times$	$\times$	83.29
4	$\surd$	$\surd$	$\surd$	$\surd$	$\times$	86.84
5	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	91.33

with the addition of skeleton frames obtains an accuracy of 83.29%, which is 2.75% more accurate than that in Group 2. This indicates that the strategy incorporating skeleton frames in addition to RGB frames helps the model recognize fine-grained activities. (iii) Compared to Group 3, the model of Group 4 adds the strategy of using the same feature extractor for RGB and skeleton frames. The accuracy of Group 4 is improved by 3.55% over Group 3. This suggests that using the same feature extractor ensures that the extracted features are consistent across the two visual modalities (RGB and skeleton frames). (iv) Compared with Group 4, the model of Group 5 adds the strategy of using 50%T prediction, and the accuracy increases by 4.49% compared to Group 4. This indicates that the 50%T prediction model takes into account the 50% beginning activity in the continuous fine-grained activity classification, which increases the accuracy of handling transitions between adjacent activities. Overall, compared to Group 1 ~ Group 4, Group 5 (the proposed model) obtains a higher accuracy of 91.33%, which is a significant improvement over the other models.

The sample in Fig. 17 shows the ground truth and recognition results of continuous fine-grained activities with an input length of 20 frames and a recognition frequency of 6 Hz. The camera frame rate is 30 fps. The recognition starts from the 20th action frame. We use different colored bars to indicate the different activity labels of different frames. We find that: (i) Our model correctly recognizes more than 95% of the activities and detects activities that change continuously in a short period, e.g., between around the 1100th and 1500th frames. The only incorrect recognition occurs around the 1700th frame when A10 is recognized as A1, but our model quickly finds the correct label and corrects itself within 10 frames, i.e., < 0.33 s. (ii) Although there is a small delay in identifying ambiguous action frames between two

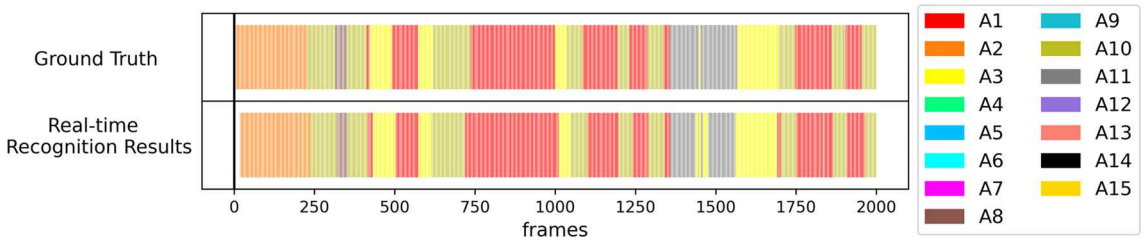


Fig. 17 Ground truth and recognition results of continuous assembly activities

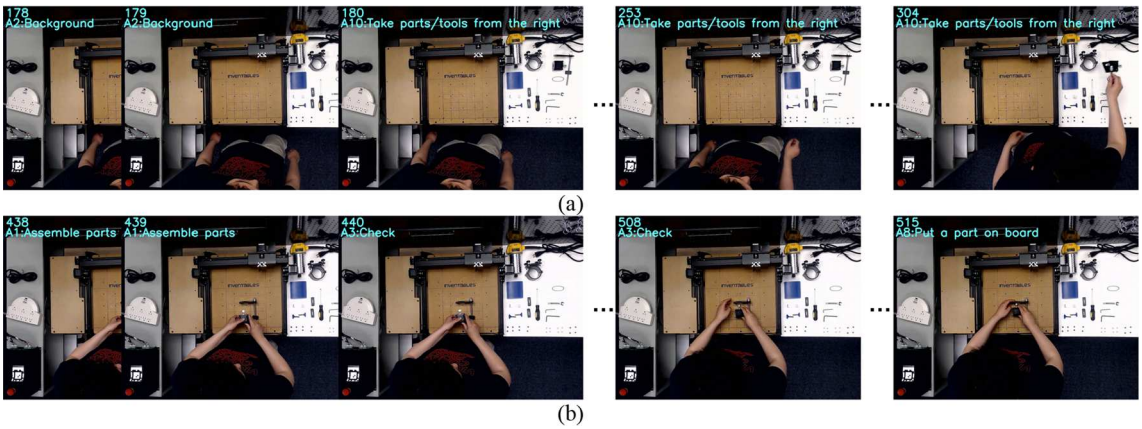


Fig. 18 Samples of continuous fine-grained activity recognition

adjacent activities, e.g., around the 1400th frame time, no actual action in the ground truth is missing in the continuous recognition results. (iii) Our model can sensitively perceive the upcoming activity changes with a time lag of about 5 frames, i.e., < 0.16 s. This temporal delay occurs mainly in identifying ambiguous action frames between two adjacent activities, e.g., at the 500th frame time and 1050th frame time. (iv) Our model has an accurate prediction capability, e.g., at about 600th and 730th frame time, our model correctly predicts the upcoming activities about 10 frames ahead.

Some samples of continuous fine-grained activity recognition are shown in Fig. 18. We output the frame time and activity recognition results in the upper left corner of the frames. The worker in Fig. 18a is standing in front of the work platform and preparing to take a part from the right platform. Our model can accurately predict that the worker is about to perform the activity A10, i.e., *take parts/tools from right*, before the worker fully extends his hand to reach a part on the right platform. In Fig. 18b, the worker randomly checks (A3) the mounting orientation and angle of the parts during the assembly (A1). Although the two activities A1 (*assemble parts*) and A3 (*check*) are very similar and the duration of the check (A3) activity is very short, our model accurately and timely detects the activity A3 (*check*) within 4 frames, i.e., < 0.13 s. The above results demonstrate our model's robustness, sensitivity, and accuracy. A video of the demonstration is available. (link: https://youtu.be/-clVPg2jHdc).

Comparison with the state-of-the-art methods

To validate the generalization ability of our model, we apply our model to a commonly used public dataset, UCF101 (Soomro et al., 2012b), which is an activity recognition dataset of real action videos collected from YouTube, with 13,320 videos from 101 activity categories. The UCF101 gives the largest diversity in activities and large variations in camera motion, object appearance, object scale, viewpoint, cluttered background, etc. The performance comparison of several state-of-the-art models using RGB frames and pre-trained on the ImageNet dataset with our model on the UCF101 dataset are presented in Table 10. Our model achieves the highest recognition accuracy compared with several other methods in the literature.

We compare our recognition model with the state-of-the-art models on the recognition of continuous fine-grained activities (containing ambiguous activity frames). The comparisons are given in Table 11, in which we applied the models in Lea et al. (2017), Simonyan and Zisserman (2014b) and Ma et al. (2021) to our dataset. The results in Table 11 show that our model has the highest accuracy of 91.33%, which is 6.83%, 4.13%, and 3.21% higher than the results obtained using the models in Lea et al. (2017), Simonyan and Zisserman (2014b) and Ma et al. (2021), respectively. These

Table 10 Accuracy (%) comparison of our model with existing models on the UCF101 public dataset

Method	Accuracy	Precision	Recall	F_1 Score
(Zhu et al., 2020)	96.90	94.99	82.67	88.40
(Huang & Bors, 2022)	97.60	96.48	87.26	91.64
(Crasto et al., 2019)	97.80	96.02	85.90	90.68
(Stroud et al., 2020)	97.90	96.11	86.08	90.82
(Carreira & Zisserman, 2017)	98.00	96.08	85.92	90.72
(Qiu et al., 2019)	98.20	97.23	89.77	93.35
Our model	98.48	97.68	91.23	94.34

Table 11 Performance (%) comparison of our model with existing models on assembly recognition

Method	Accuracy	Precision	Recall	F_1 Score
(Lea et al., 2017)	84.50	79.21	48.79	60.39
(Simonyan & Zisserman, 2014b)	87.20	83.17	55.26	66.40
(Ma et al., 2021)	88.12	84.54	57.71	68.60
Our model	91.33	85.71	59.98	70.57

comparison results show that our model achieves the highest recognition accuracy of continuous fine-grained assembly activities compared with the other methods in the literature.

We used the t-SNE (t-Distributed Stochastic Neighbor Embedding) method (Ullah et al., 2022) to visually compare the performance of our model with that in Lea et al. (2017); Simonyan & Zisserman, (2014b); Ma et al., (2021) on our dataset. The results are shown in Fig. 19a–d, where the high-dimensional feature data are visualized in two dimensions, providing insight into the underlying structure of the feature embeddings. We found that the clusters of our model are more clearly defined and grouped compared to other models. To provide a clearer visualization, we opted to zoom in on the clusters by using smaller scales in Fig. 19d. In addition, we use silhouette scores to measure the similarity of data points within clusters to those in other clusters. Our model yields a higher silhouette score of 0.009, indicating that the clusters are more clearly defined and separated in the t-SNE visualization. Thus, visual and quantitative analyses show that our model performs better than other models.

Conclusion

In this paper, we create a fine-grained activity dataset that contains 15 fine-grained activities in the assembly of a desktop carving machine. We design a new two-stage network to classify the fine-grained activities in assembly. In this proposed network, we fuse multi-visual modalities, specifically red–green–blue (RGB) and hand skeleton frames, to capture fine-grained activity details. In the first stage, we use a pre-trained VGG-16 model to extract the scene-level activity features. In the second stage, this network uses the Long Short-Term Memory (LSTM) to extract the activities' temporal-level features. In designing this network, we compare the effects of different data input lengths, different types of pre-trained models in transfer learning, different Recurrent Neural Networks (RNNs), and different fusion mechanisms on recognition performance. We conduct fine-grained activity prediction using the partial video observation method and propose a new fusion recognition-prediction model to recognize and predict continuous fine-grained activities. The experimental results using the trimmed videos as the inputs show that: (i) An average recognition accuracy of 99.98% is obtained using the recognition model with an input length of 20 frames, a VGG-16 pre-training model, an LSTM structure, and a late-fusion mechanism. (ii) An average prediction accuracy of 97.82% is obtained using 50% of the activity onset information in activity prediction. The experimental results using an untrimmed video with continuous fine-grained activities as the inputs show that: (i) Our fusion model achieves an average recognition accuracy of 91.33% with a speed of 0.032 s (faster than real-time) for an activity sequence with an input frame rate of 30 frames per second (fps). (ii) This fusion model correctly detects activity change within about 5 frames (< 0.16 s) and correctly predicts upcoming activities about 10 frames (0.33 s) in advance. Compared with state-of-the-art models in the literature, our fusion model outperforms those models using RGB frames and pre-trained data on the ImageNet dataset in recognizing the UCF101 public dataset. Our model also beats the state-of-the-art models in recognizing continuous fine-grained assembly activities. The comparison results show that our model has established a new annotation baseline in the recognition and prediction of continuous assembly activities.

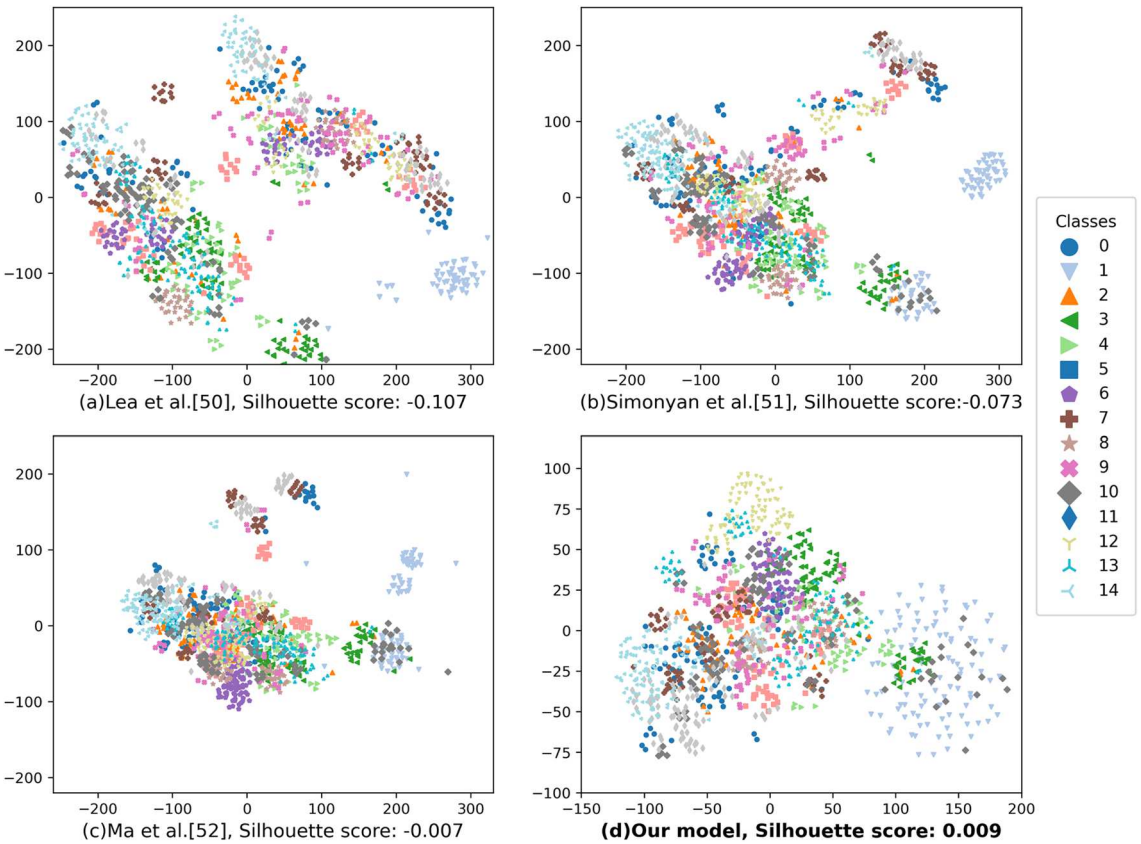


Fig. 19 Performance of different models in t-SNE visualization

Several future extensions of our study described in this paper are envisaged, including exploring other modalities, such as brain waves and eye gaze, to identify the worker's intention in assembly, improving the recognition of the continuous fine-grained activities in assembly by combing the information about tools and parts used in the assembly, and designing a real-time monitor system to improve assembly efficiency by providing assembly guidance based on the worker's current activities.

Acknowledgements This research work is financially supported by the National Science Foundation grants CMMI-1646162 and CMMI-1954548 and by the Intelligent Systems Center at Missouri University of Science and Technology. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Funding National Science Foundation, CMMI-1646162, Ming C. Leu, CMMI-1954548, Zhaozheng Yin

References

Ahn, D., Kim, S., Hong, H. and Ko, B.C., 2023. STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (pp. 3330–3339). https://doi.org/10.48550/arXiv.2210.07503

Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. *Electronics, 10*(9), 1036. https://doi.org/10.3390/electronics10091036

Al-Amin, M., Qin, R., Moniruzzaman, M., Yin, Z., Tao, W., & Leu, M. C. (2021). An individualized system of skeletal data-based CNN classifiers for action recognition in manufacturing assembly. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-021-01815-x

Byrne, J., Castañón, G., Li, Z. and Ettinger, G., 2023. Fine-grained Activities of People Worldwide. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3308–3319). https://doi.org/10.48550/arXiv.2207.05182

Carreira, J. and Zisserman, A., 2017. Quo vadis, action recognition, a new model, and the kinetics dataset. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299–6308). https://doi.org/10.1109/CVPR.2017.502

Chan, J. Y. L., Bea, K. T., Leow, S. M. H., Phoong, S. W., & Cheng, W. K. (2023). State of the art: A review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review, 56*(1), 749–780. https://doi.org/10.1007/s10462-022-10183-8

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., & Liu, Y. (2021). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR), 54*(4), 1–40. https://doi.org/10.1145/3447744

Chen, H., Leu, M. C., & Yin, Z. (2022). Real-time multi-modal human-robot collaboration using gestures and speech. *Journal of*

Manufacturing Science and Engineering. https://doi.org/10.1115/1.4054297

Chen, H., Leu, M.C., Tao, W. and Yin, Z., 2020a, November. Design of a real-time human-robot collaboration system using dynamic gestures. In ASME International Mechanical Engineering Congress and Exposition. American Society of Mechanical Engineers. Doi: https://doi.org/10.1115/IMECE2020-23650

Chen, H., Tao, W., Leu, M.C. and Yin, Z., 2020b, July. Dynamic gesture design and recognition for human-robot collaboration with convolutional neural networks. In: *International Symposium on Flexible Automation* (Vol. 83617, p. V001T09A001). American Society of Mechanical Engineers. https://doi.org/10.1115/ISFA2020-9609

Cho, K., Van Merriënboer, B., Bahdanau, D. and Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259. https://doi.org/10.48550/arXiv.1409.1259

Cho, J., Baskar, M.K., Li, R., Wiesner, M., Mallidi, S.H., Yalta, N., Karafiat, M., Watanabe, S. and Hori, T., 2018. Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling. In: *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 521–527). IEEE. https://doi.org/10.1109/SLT.2018.8639655

Crasto, N., Weinzaepfel, P., Alahari, K. and Schmid, C., 2019. Mars: Motion-augmented RGB stream for action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7882–7891). https://doi.org/10.1109/CVPR.2019.00807

Fu, Z., He, X., Wang, E., Huo, J., Huang, J., & Wu, D. (2021). Personalized human activity recognition based on integrated wearable sensor and transfer learning. *Sensors, 21*(3), 885. https://doi.org/10.3390/s21030885

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). https://doi.org/10.48550/arXiv.1512.03385

https://www.tensorflow.org/guide/keras/masking_and_padding

Hu, Z., Yu, T., Zhang, Y. and Pan, S., 2020, September. Fine-grained activities recognition with coarse-grained labeled multi-modal data. In: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (pp. 644–649). https://doi.org/10.1145/3410530.3414320

Huang, G. and Bors, A.G., 2022. Busy-Quiet Video Disentangling for Video Classification. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1341–1350). https://doi.org/10.48550/arXiv.2103.15584

Jones, J. D., Cortesa, C., Shelton, A., Landau, B., Khudanpur, S., & Hager, G. D. (2021). Fine-grained activity recognition for assembly videos. *IEEE Robotics and Automation Letters, 6*(2), 3728–3735. https://doi.org/10.1109/LRA.2021.3064149

Kapidis, G., Ronald P., Elsbeth V. D., Lucas Noldus, and Remco Veltkamp. "Egocentric hand track and object-based human action recognition." In 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 922–929. IEEE, 2019. https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00185

Khan, M. A., Akram, T., Zhang, Y. D., & Sharif, M. (2021). Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework. *Pattern Recognition Letters, 143*, 58–66. https://doi.org/10.1016/j.patrec.2020.12.015

Kobayashi, T., Aoki, Y., Shimizu, S., Kusano, K. and Okumura, S., 2019, November. Fine-grained action recognition in assembly work scenes by drawing attention to the hands. In: *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 440–446). IEEE.

Kumar, Y., & Gupta, S. (2023). Deep transfer learning approaches to predict glaucoma, cataract, choroidal neovascularization, diabetic macular edema, drusen and healthy eyes: An experimental review. *Archives of Computational Methods in Engineering, 30*(1), 521–541. https://doi.org/10.1007/s11831-022-09807-7

Lea, C., Flynn, M.D., Vidal, R., Reiter, A. and Hager, G.D., 2017. Temporal convolutional networks for action segmentation and detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 156–165). https://doi.org/10.1109/CVPR.2017.113

Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B. and Wang, L., 2020. Tea: Temporal excitation and aggregation for action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 909–918). https://doi.org/10.48550/arXiv.2004.01398

Ma, C. Y., Chen, M. H., Kira, Z., & AlRegib, G. (2021). TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication, 71*, 76–87. https://doi.org/10.1016/j.image.2018.09.003

Marszalek, M., Laptev, I. and Schmid, C., 2009. Actions in context. In: *2009 IEEE conference on computer vision and pattern recognition* (pp. 2929–2936). IEEE. https://doi.org/10.1109/CVPR.2009.5206557

Mekruksavanich, S., & Jitpattanakul, A. (2022). Multimodal wearable sensing for sport-related activity recognition using deep learning networks. *Journal of Advances in Information Technology*. https://doi.org/10.12720/jait.13.2.132-138

Pan, S., Berges, M., Rodakowski, J., Zhang, P., & Noh, H. Y. (2020). Fine-grained activity of daily living (ADL) recognition through heterogeneous sensing systems with complementary spatiotemporal characteristics. *Frontiers in Built Environment*. https://doi.org/10.3389/fbuil.2020.560497

Qiu, Z., Yao, T., Ngo, C.W., Tian, X. and Mei, T., 2019. Learning Spatio-temporal representation with local and global diffusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12056–12065). https://doi.org/10.48550/arXiv.1906.05571

Rohrbach, M., Amin, S., Andriluka, M. and Schiele, B., 2012, June. A database for fine-grained activity detection of cooking activities. In: *2012 IEEE conference on computer vision and pattern recognition* (pp. 1194–1201). IEEE. https://doi.org/10.1109/CVPR.2012.6247801

Rude, D. J., Adams, S., & Beling, P. A. (2018). Task recognition from joint tracking data in an operational manufacturing cell. *Journal of Intelligent Manufacturing, 29*(6), 1203–1217. https://doi.org/10.1007/s10845-015-1168-8

Ryoo, M.S. and Aggarwal, J.K., 2009, September. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: *2009 IEEE 12th international conference on computer vision* (pp. 1593–1600). *IEEE*. https://doi.org/10.1109/ICCV.2009.5459361

Schuldt, C., Laptev, I. and Caputo, B., 2004, August. Recognizing human actions: a local SVM approach. In: *Proceedings of the 17th International Conference on Pattern Recognition*. ICPR 2004. (Vol. 3, pp. 32–36). IEEE. https://doi.org/10.1109/ICPR.2004.1334462

Sherafat, B., Ahn, C. R., Akhavian, R., Behzadan, A. H., Golparvar-Fard, M., Kim, H., Lee, Y. C., Rashidi, A., & Azar, E. R. (2020). Automated methods for activity recognition of construction workers and equipment: State-of-the-art review. *Journal of Construction Engineering and Management, 146*(6), 03120002.

Simonyan, K. and Zisserman, A., 2014a. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Simonyan, K. and Zisserman, A., 2014b. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems. https://doi.org/10.5555/2968826.2968890

Singh, B., Marks, T.K., Jones, M., Tuzel, O. and Shao, M., 2016. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1961–1970). https://doi.org/10.1109/CVPR.2016.216

Soomro, K., Zamir, A.R. and Shah, M., 2012a. UCF101: A dataset of 101 human action classes from videos in the wild. arXiv preprint arXiv:1212.0402. https://doi.org/10.48550/arXiv.1212.0402

Soomro, K., Zamir, A.R. and Shah, M., 2012b. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402. https://doi.org/10.48550/arXiv.1212.0402

Stroud, J., Ross, D., Sun, C., Deng, J. and Sukthankar, R., 2020. D3d: Distilled 3d networks for video action recognition. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 625–634). https://doi.org/10.48550/arXiv.1812.08249

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826). https://doi.org/10.48550/arXiv.1512.00567

Tao, W., Al-Amin, M., Chen, H., Leu, M. C., Yin, Z., & Qin, R. (2020). Real-time assembly operation recognition with fog computing and transfer learning for human-centered intelligent manufacturing. *Procedia Manufacturing, 48*, 926–931. https://doi.org/10.1016/j.promfg.2020.05.131

Tian, C., Xu, Y., & Zuo, W. (2020). Image denoising using deep CNN with batch renormalization. *Neural Networks, 121*, 461–473. https://doi.org/10.1016/j.neunet.2019.08.022

Ullah, B., Kamran, M., & Rui, Y. (2022). Predictive modeling of short-term rockburst for the stability of subsurface structures using machine learning approaches: T-SNE. *K-Means Clustering and XGBoost. Mathematics, 10*(3), 449. https://doi.org/10.3390/math10030449

Xia, L., Chen, C.C. and Aggarwal, J.K., 2012, June. View invariant human action recognition using histograms of 3d joints. In: *2012 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 20–27). *IEEE*. https://doi.org/10.1109/CVPRW.2012.6239233

Xiao, J., Jing, L., Zhang, L., He, J., She, Q., Zhou, Z., Yuille, A. and Li, Y., 2022. Learning from temporal gradient for semi-supervised action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3252–3262. https://doi.org/10.48550/arXiv.2111.13241

Yao, B., Khosla, A. and Fei-Fei, L., 2011. Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. a) A, 1(D2), p.D3. https://www.semanticscholar.org/paper/Classifying-Actions-and-Measuring-Action-Similarity-Yao-Khosla/9612fd66fcd3902bc267a62c146398eb8d30830e

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation, 31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Zhang, C., Zou, Y., Chen, G. and Gan, L., 2020a. Pan: Towards fast action recognition via learning persistence of appearance. arXiv preprint arXiv:2008.03462. https://arxiv.org/abs/2008.03462

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L. and Grundmann, M., 2020b. Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214. https://doi.org/10.48550/arXiv.2006.10214

Zheng, T., Ardolino, M., Bacchetti, A., & Perona, M. (2021). The applications of Industry 4.0 technologies in manufacturing context: a systematic literature review. *International Journal of Production Research, 59*(6), 1922–1954. https://doi.org/10.1080/00207543.2020.1824085

Zhu, L., Tran, D., Sevilla-Lara, L., Yang, Y., Feiszli, M. and Wang, H., 2020, April. Faster recurrent networks for efficient video classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 13098–13105). https://doi.org/10.1609/aaai.v34i07.7012