International Journal of Computer Vision and Signal Processing, 14(1), 9-17(2024) ORIGINAL ARTICLE

Hybrid Deep Learning for Assembly Action Recognition in Smart Manufacturing

Abdul Matin*

School of Computer Science, University of Technology Sydney, Australia

Md Rafiqul Islam

Data Science Institute, University of Technology Sydney, Australia

Yeqian Zhu, Xianzhi Wang, Huan Huo

School of Computer Science, University of Technology Sydney, Australia

Guandong Xu

Data Science Institute, University of Technology Sydney, Australia

Abstract

Deep learning algorithms have become essential in assembly action recognition (AAR) for driving advancements in intelligent manufacturing. While numerous sensor systems and algorithms are developing, their real-world applicability and robustness within the manufacturing sector need validation. Artificial intelligence (AI) applications in manufacturing have gained significant attraction in both academic and industrial circles. One key aspect of future smart manufacturing is identifying the actions of manufacturing workers, particularly monitoring repetitive assembly tasks, to guide them and improve efficiency. This recognition facilitates real-time efficiency measurement and evaluation of workers while providing augmented reality instructions to enhance their performance on the job. This paper introduces a hybrid deep-learning approach combining 3D CNN and ConvLSTM2D models to monitor assembly tasks to recognize human actions within the manufacturing context. The model's performance is evaluated through simulations conducted on the HA4M dataset, comprising diverse multimodal data-capturing actions executed by various individuals constructing an epicyclic gear train (EGT). The proposed hybrid model demonstrated superior performance on the HA4M dataset relative to baselines.

Contribution of the Paper: A novel hybrid deep learning model that outperforms multiple state-of-the-art models for assembly action recognition.

Keywords: Deep Learning, Human Activity Recognition, Convolutional Neural Networks, Smart Manufacturing

ISSN: 2186-1390 (Online)
http://cennser.org/IJCVSP

Article History:
Received: 12 September 2023
Revised: 26 September 2023
Accepted: 10 January 2024
Published Online: 30 January 2024

1. INTRODUCTION

Assembly action recognition in manufacturing is the process of automatically identifying and classifying the actions of workers during the assembly process. This can be done using a variety of sensors, such as cameras, depth sensors, and inertial measurement units (IMUs). The goal of assembly action recognition is to improve manufacturing efficiency and quality by providing real-time feedback to workers, identifying errors, and tracking productivity. Recently, there has been growing interest in research on assembly task monitoring and workers' activity recognition that can potentially improve human-robot collaboration, task efficiency, and real-time instruction provision using augmented reality. Workers are involved in opera-

*Corresponding author

Email addresses: abdul.matin@student.uts.edu.au (Abdul Matin), mdrafiqul.islam@uts.edu.au (Md Rafiqul Islam), yeqian.zhu@student.uts.edu.au (Yeqian Zhu), xianzhi.wang@uts.edu.au (Xianzhi Wang), huan.huo@uts.edu.au (Huan Huo), guandong.xu@uts.edu.au (Guandong Xu)

tions such as assembling manufacturing products that are mostly harmonious and symmetric [1]. Manufacturing industries increasingly use industrial robots to improve efficiency and reduce risks for human operators [2]. Different communication protocols, including wired, wireless, and remote approaches, have been developed to increase manufacturing output. Some researchers have even developed remote-controlled robots that users can operate from a distance [3, 4]. However, comprehensively monitoring all aspects of the manufacturing environment is a complex phenomenon. Recognizing the activities of manufacturing workers requires a deep understanding of context and the ability to track objects in real time. This task is difficult in industrial assembly settings, which are constantly changing [1, 5]. As a result, identifying different activities becomes even more challenging when tracking workers' movements and visually discerning their actions. This research highlights the need for more research on monitoring the activities of manufacturing workers.

Therefore, the main goal of this study is to develop a deep learning-based smart manufacturing workers' action recognition system that can recognize the actions of assembly tasks. This paper presents three outlines for assembly action recognition in manufacturing.

- First, we explore the performances of four baseline deep learning algorithms on the HA4M dataset. These algorithms are convLSTM, LSTM with VG16, LRCN, and 3D CNN.
- Second, we design a hybrid model combining ConvLSTM and 3D CNN. In this model, we use the output of ConvLSTM as input of 3D CNN. We also simplify the model to reduce its complexity and improve its performance.
- Third, we propose a hybrid model using 3D CNN and ConvLSTM. In this model, the output of the 3D CNN layer is used as input of the ConvLSTM layer. The proposed hybrid model outperforms the baseline models.

The proposed hybrid assembly action recognition approach can revolutionize manufacturing by enhancing workers' efficiency, flexibility, and sustainability, all of which are the central objectives of Industry 4.0. Their practical implementation in manufacturing settings aligns with the vision of a more intelligent, connected, and agile manufacturing industry. The remainder of this paper is organized as follows: Section 2 presents the related literature on assembly action recognition; Section 3 provides an elaborated description of the HA4M dataset and proposed approach to developing the hybrid model; Section 4 and Section 5 assesses the training and testing performance of the baseline models and proposed hybrid models; Finally, Section 6 summarizes the conclusions of the study and outline the future research directions.

2. RELATED WORKS

Deep learning and machine learning are now essential tools for identifying human actions. This field has grown in popularity, as evidenced by the increasing amount of research on the topic. The development of observation systems for various applications, such as video surveillance, safety, smart home security, ambient assisted living, and healthcare, has been driven by technological advancements such as the availability of low-cost sensors and video camera-based systems. These systems can detect and analyze human actions, providing valuable information to improve safety, security, and well-being. Despite this progress, there still needs to be more research on assembly action recognition [5, 6, 7, 8] for sustainable manufacturing [9], and the lack of public datasets is hindering the development and comparison of new methods [1]. In computer vision, deep learning methods have been extensively researched, and recent advancements have made it possible to detect human activity in images and videos with exceptional accuracy [10]. However, visual-based recognition accuracy can be reduced by occlusion.

The study of action recognition within manufacturing assembly is an emerging research area and encompasses various innovative methodologies to date, notably hybrid models, deep learning frameworks, multimodal sensor integration, and computer vision techniques. One such hybrid model [11] adeptly merges a convolutional neural network (CNN) and variable-length Markov modeling (VMM), achieving 94.7% accuracy in action recognition and effective collaboration context comprehension. Rigorous assessments within simulated assembly environments validate its robust accuracy. Ultrasonic and IMU sensor-based assembly action recognition systems [12] introduce for the maintenance of bicycles using the Hidden Markov Model classifier, where an unsupervised measurement method utilizes to estimate lead time for factory work using signals from a smartwatch with an IMU sensor. Another investigation employed numerous IMU sensors to partition and categorize worker actions in car manufacturing operations. Additionally, a wrist-worn IMU sensor was employed to record arm motions and categorize five distinct activities within the setting of industrial assembly lines [13]. Moreover, [6] employs deep learning via YOLOv3 and CPM algorithms to discern recurring assembly actions and anticipate operation durations. Notably, action recognition achieves 92.8% accuracy, while operating time estimation attains 82.1%. Additionally, Al-Amin et al. [14] introduce a method for assembly action recognition by fusing data from IMUs, EMG sensors, and vision sensors, yielding a notable 92.8% accuracy—similarly, another research introduced by this author in [5] using deep learning with CNNs tailored to individual workers, achieving an impressive 94% accuracy in recognizing assembly tasks.

More recently, deep learning techniques have been introduced to recognize worker activity in studies on human-robot collaboration [15]. In the context of human-robot

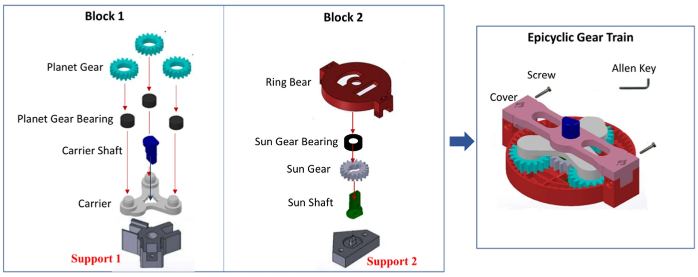

Figure 1: Three distinct phases of assembling an Epicyclic Gear Train (EGT) [1]

collaboration, [16] highlights Hidden Markov Models (HMMs) efficacy, tracking worker movements via camera data to enhance action recognition. Intriguingly, layered HMMs model basic movements while calculated trajectories optimize efficiency. Duarte Moutinho et al. [17] present a system using ResNet and LSTM networks to comprehend human actions in collaborative assembly, attaining 96.65% accuracy. Transfer learning is featured in [18], leveraging a pre-trained Kinetics model for human-robot collaborative assembly, resulting in a 92.8% accuracy post-fine-tuning. Furthermore, a human-robot collaborative assembly system proposed by H. Goto et al. [19] employs real-time human action recognition, enhancing assembly efficiency and safety through FSM task models and vision-based recognition. An assembly-plan-from-observation (APO) technique [20] captures fundamental task representations by recovering assembly relations from human tasks, enabling the generation of robot programs. Similarly, using graph and temporal convolution networks, a spatiotemporal-based approach [21] significantly enhances action segmentation accuracy. Lastly, a novel assembly action recognition architecture [22] involving multi-camera setups and LSTM networks outperforms previous methods, meticulously analyzing assembly actions and human-object relationships. However, [23]'s computer vision algorithm for tracking manual assembly tasks faces limitations in accommodating real-world illumination variations.

3. METHODOLOGY

This section describes the HA4M [1] dataset and deep learning techniques used for assembly action recognition in manufacturing. Subsections 3.1 and 3.2 introduce the dataset properties and facilitate the methodological analysis.

3.1. Dataset Description

In this study, we used RGB data of the HA4M [1] dataset for simulating models. This dataset consists of 217 videos capturing the process of assembling an Epicyclic Gear Train (EGT) by 41 participants, comprising 15 females and 26 males. The dataset was meticulously crafted to evaluate the performance of individuals with varying skill levels, encompassing intricate tasks executed diversely by individuals of differing ages and competencies. It offers a diverse range of multi-modal data capturing the activities associated with EGT assembly in a controlled laboratory environment, featuring six distinct types of data: RGB images, Depth maps, IR images, RGB-to-Depth-Aligned images, Point Clouds, and Skeleton data. Precisely, 41 subjects undertook numerous assembly attempts, performing 12 distinct actions. The data collection process employed a Microsoft Azure Kinect, which integrates an RGB camera, a depth camera, and InfraRed (IR) emitters. The HA4M dataset is an invaluable resource for researchers engaged in developing and assessing assembly action recognition systems. It furnishes a realistic and challenging dataset that serves as a robust benchmark for evaluating the efficacy of various methodologies.

3.1.1. Overview of assembling an Epicyclic Gear Train (EGT)

The process of assembling an Epicyclic Gear Train (EGT) comprises three distinct phases, as depicted in figure 1. The initial stages involve the separate assembly of Blocks 1 and 2, followed by the eventual integration of these two blocks. The configuration of the EGT entails a total of 13 constituent components, encompassing eight elements for the construction of Block 1, four elements for Block 2, and a cover responsible for linking Blocks 1 and 2. Upon completing the individual blocks, the secure attachment of both blocks is achieved by using two screws and an Allen key, thereby finalizing the EGT assembly process. Furthermore, figure 1 provides a visual representation of the incorporation of two supports strategically designed to facilitate the assembly of each respective block.

3.1.2. Actions description

The HA4M dataset consists of 12 distinct actions divided into three phases: the first four actions are for making Block 1; the following four actions are for creating Block 2, and the final four actions involve putting the two blocks together to complete the EGT.

- **Phase 1:** The initial action entails picking up and placing the Carrier, denoted as action 1. Subsequently, action 2 involves picking up three Gear Bearings one by one and arranging them onto the Carrier. It is followed by action 3, which consists of picking up and placing down three Planet Gears. Concluding the assembly of Block 1 is action 4, where the Carrier Shaft is picked up and placed.

- **Phase 2:** Pick up and place the Sun Shaft in its place, denoted as action 5. After that, pick up and put down the Sun Gear, action 6. Then, pick up and set down the Sun Gear Bearing, action 7. To complete Block 2, pick up and place the Ring Bear, action 8.

- **Phase 3:** Pick up Block 2 and put it on Block 1. This is action 9. Then, lift and put down the Cover, which is action 10. After that, pick up and place two Screws, which is step 11. Finally, pick up the Allen Key, turn both screws, put the Allen Key back, and finish assembling the EGT.

3.2. Methodological Analysis

We employed four baseline models for recognizing manufacturing assembly actions on the HA4M dataset: ConvLSTM2D, LSTM with VG16, LRCN, and 3D CNN. Later we designed two hybrid models using the baseline models. We combined ConvLSTM2D and 3D CNN for the first hybrid model and then simplified it to reduce the complexity of the model and improve performance. In the second hybrid model, we combined 3D CNN and ConvLSTM2D. The baseline model and the hybrid model descriptions are as follows:

3.2.1. ConvLSTM2D

The model's architecture comprises four ConvLSTM2D layers, succeeded by a flatten layer, and ultimately a dense layer featuring softmax activation. The initial ConvLSTM2D layer employs four filters and a (3, 3) kernel size and applies the tanh activation function. A recurrent_dropout parameter of 0.2 is employed, and the return_sequences parameter is set to True, aligning with the ConvLSTM2D layer's recurrent nature designed for sequential data processing. In the subsequent ConvLSTM2D layers, specifically the second, third, and fourth ones, filter counts of 8, 16, and 32 are employed correspondingly. MaxPooling3D layers are incorporated after each ConvLSTM2D layer, halving the output tensor's dimensions to control the model's complexity and prevent overfitting. The model integrates TimeDistributed layers, a key element enabling the application of ConvLSTM2D operations to every frame of the assembly action. Meanwhile, Dropout layers are thoughtfully integrated within the ConvLSTM2D layers to stochastically deactivate certain neurons, thus safeguarding against overfitting the training and testing dataset. The Dense layer within the model undertakes the classification of assembly actions into one of the 12 predefined categories. The softmax activation function is adopted in conjunction with this dense layer.

3.2.2. LSTM with VG16

Initially, the model employs a pre-trained VGG16 architecture to extract features from assembly action video frames. This VGG16 model is applied in a TimeDistributed manner, enabling individual processing of each frame. Subsequently, the outcome of the VGG16 model is passed through a Flatten layer, converting the output tensor into a flattened vector. This vector is subsequently forwarded to an LSTM layer with 128 neurons. The LSTM layer is succeeded by a Dropout layer, strategically introduced to avert potential overfitting of the LSTM layer on the training data. Following the LSTM stage, the resultant output advances to a Dense layer boasting a dozen neurons. The output of the LSTM layer is then passed to a Dense layer with 12 neurons. The softmax activation function is used in the last Dense layer.

3.2.3. LRCN

The model comprises four Conv2D layers with (3, 3) kernels, padding as same, ReLU activation, and filter counts: 16, 32, 64, and 64, respectively. Following each Conv2D layer, a MaxPooling2D layer with (4, 4) pooling and a Dropout layer dropping 25% of neurons are incorporated. A Flatten layer follows, transforming the output tensor into a flat vector. Subsequently, the Flatten layer's output is directed to an LSTM layer with 32 neurons. This LSTM output, in turn, advances to a Dense layer of 12 neurons, employing softmax activation for the final classification outcomes.

3.2.4. 3D CNN

The 3D convolutional neural network (3D CNN) architecture comprises a pair of Conv3D layers, each defined by a kernel size of (3, 3, 3) and employing the ReLU activation function. The initial Conv3D layer is configured with 32 filters, while the subsequent layer uses 64 filters, enhancing the network's feature extraction capacity. After each Conv3D layer, a singular MaxPooling3D layer, characterized by a pool size of (2, 2, 2), is incorporated to achieve effective spatial dimension reduction. The fifth layer in the network is the Flatten layer, strategically employed to transform the output tensor into a flattened vector form. For assembly action classification into one of the predefined 12 categories, the architecture incorporates two Dense layers featuring 128 and 12 neurons, respectively. The softmax activation function is employed within the terminal Dense layer, ensuring the model's output conforms to a well-structured probability distribution.

3.2.5. ConvLSTM2D+3D CNN

This hybrid model in figure 2 merges ConvLSTM2D and 3D CNN techniques, intertwining their strengths. The model combines two ConvLSTM2D layers, two 3D CNN layers, four MaxPooling3D layers, a Flatten layer, and two dense layers. The second ConvLSTM2D output serves as input for the inaugural 3D CNN layer.

The first ConvLSTM2D layer applies four filters with a (3, 3) kernel, employs tanh activation and recurrent_dropout set 0.2, and return_sequences is True. A MaxPooling3D layer (pool_size: 1x2x2) follows, curtailing complexity. Incorporated Dropout layers within ConvLSTM2D randomly deactivate neurons. The second ConvLSTM2D layer mirrors the first, utilizing eight filters. Employing MaxPooling3D (pool_size: 1x2x2) and Dropout (20%) comes next. Subsequently, Conv3D layers follow, each with a (3, 3, 3) kernel and ReLU activation. The first layer incorporates 32 filters; the next 64 enhance feature extraction. After each Conv3D, MaxPooling3D (pool_size: 2x2x2) ensures compactness. The fifth layer is the Flatten layer. For assembly action classification (12 categories), two Dense layers (128

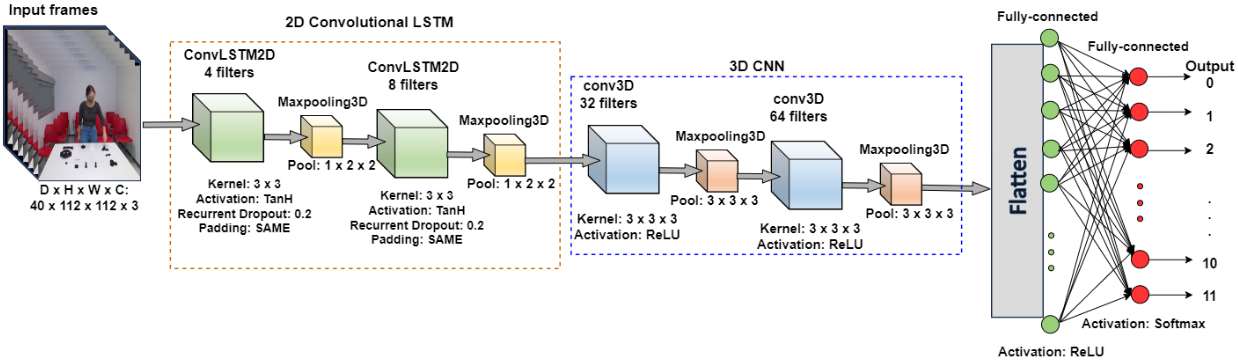

Figure 2: Hybrid deep learning model combining ConvLSTM2D and 3D CNN

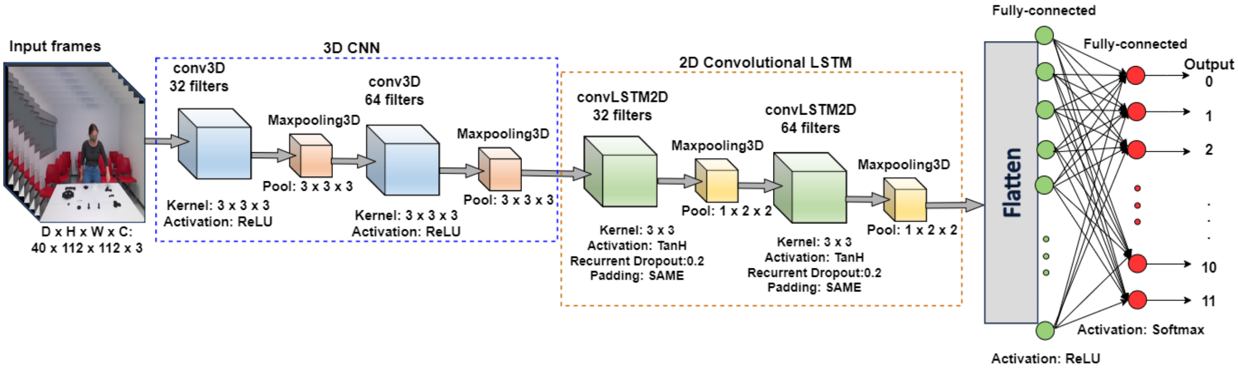

Figure 3: Hybrid deep learning model combining 3D CNN and ConvLSTM2D

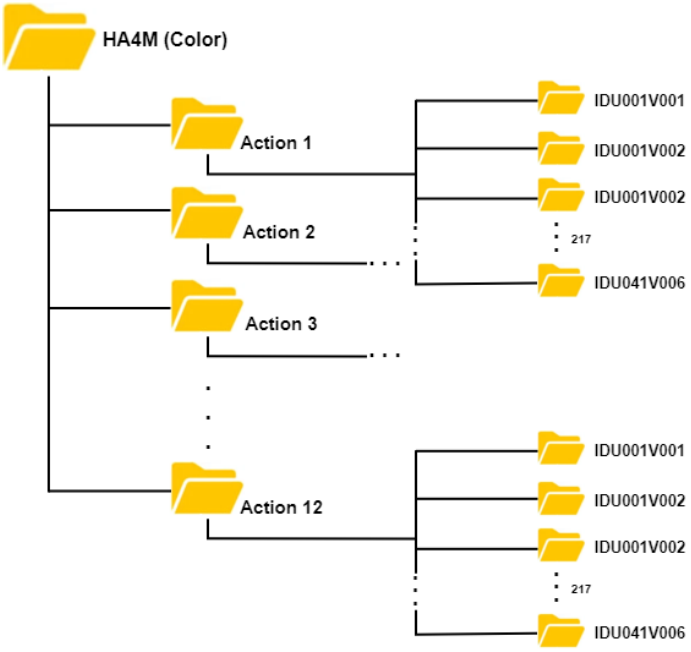

Figure 4: Dataset directory for training and testing: 12 distinct actions, each action folder contain 217 videos (sequence of RGB frames) to perform assembly task by 41 subjects

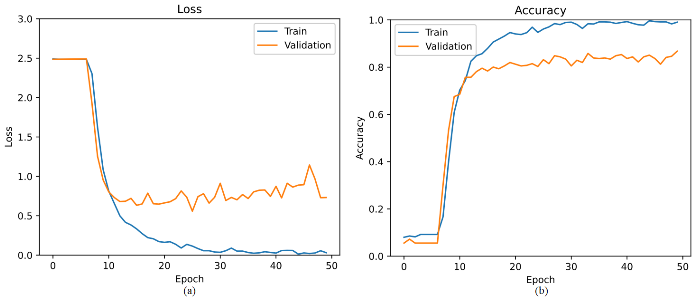

Figure 5: Training and validation history of ConvLSTM2D + 3D CNN: (a) Training loss vs Validation loss, (b) Training accuracy vs Validation accuracy

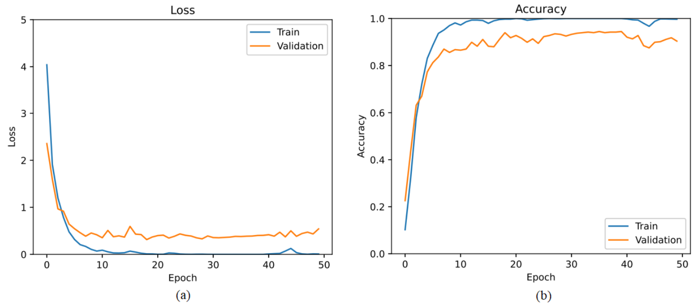

Figure 6: Training and validation history of simplified ConvLSTM2D + 3D CNN: (a) Training loss vs Validation loss, (b) Training accuracy vs Validation accuracy

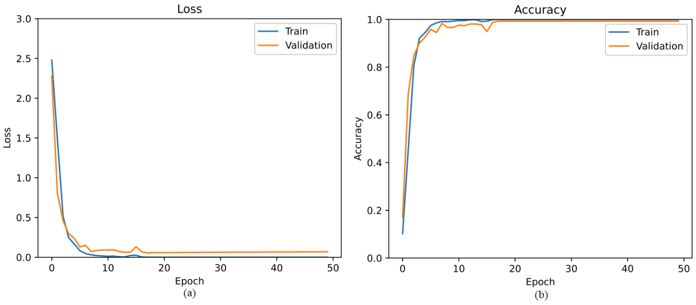

Figure 7: Training and validation history of 3D CNN + ConvLSTM2D: (a) Training loss vs Validation loss, (b) Training accuracy vs Validation accuracy

Table 1: Model training parameters description

Parameters	Value
Input shape (Number of frames, Frame Height, Frame Width, Number of channels)	40, 112, 112, 3
Batch size	32
Epochs	50
Optimizer	Adam

and 12 neurons) are implemented. Softmax activation in the terminal Dense layer guarantees a coherent probability distribution.

3.2.6. Simplified ConvLSTM2D+3D CNN

To simplify the ConvLSTM2D+3D CNN model, we excluded the second convolutional LSTM layer with max pooling 3D from the 2D convolutional LSTM part and the second convolutional layer with Maxpooling3D from the 3D CNN segment. We tried to reduce the model's complexity and make it more computationally efficient.

3.2.7. 3D CNN+ConvLSTM2D

The hybrid model in figure 3 combines the performance strengths of 3D CNNs and ConvLSTM2Ds. It has two 3D CNN layers, two ConvLSTM2D layers, four MaxPooling3D layers, a Flatten layer, and two Dense layers. The output of the second 3D CNN layer is used as input to the first ConvLSTM2D layer.

Within the Conv3D layers, a (3, 3, 3) kernel and ReLU activation are employed. The first layer integrates 32 filters, while the subsequent enhances feature extraction with 64 filters. After each Conv3D operation, MaxPooling3D (pool_size: 2x2x2) is utilized for compactness. Subsequently, the first ConvLSTM2D layer uses a (3, 3) kernel, applies tanh activation, and incorporates a recurrent_dropout rate of 0.2, while return_sequences is set to True. A MaxPooling3D layer (pool_size: 1x2x2) follows, managing intricacy. Within ConvLSTM2D, Dropout layers are strategically placed to deactivate neurons probabilistically. The second ConvLSTM2D layer employs eight filters, succeeded by MaxPooling3D (pool_size: 1x2x2) and Dropout (20%). The architecture continues with a Flatten layer. For classifying assembly actions into 12 categories, two Dense layers (128 and 12 neurons) are employed. The terminal Dense layer's softmax activation ensures a coherent and structured probability distribution.

4. PERFORMANCE EVALUATION

To train and test the deep learning models described in section 3.2, we used RGB data from the HA4M dataset. We split the dataset into the following directories, as shown in figure 4. We split the entire dataset into three segments before training the models: training, validation, and test, where 80% of the dataset was used for training and validation. The remaining 20% of the dataset was used for final testing. Training parameters for evaluated baseline and hybrid models are illustrated in Table 1.

The test performance is shown in Table 2. However, the 3D CNN model achieved the best performance (95.57%) compared to all baselines, followed by the ConvLSTM2D model, the LRCN model, and the LSTM with the VG16 model.

The training progress of the hybrid models, namely ConvLSTM2D + 3D CNN, simplified ConvLSTM2D + 3D CNN and 3D CNN + ConvLSTM2D, are depicted in figures 5, 6, and 7, respectively. The initial hybrid model did not perform satisfactorily compared to its baseline models. However, significant improvement was observed in its simplified variant, elevating the accuracy from 87.30% to 94.23%. The hybrid model employing 3D CNN and ConvLSTM2D yielded an outstanding test performance, attaining an accuracy of 99.42%.

5. Results and Discussion

We evaluated the performance of four baselines and two hybrid models for recognizing manufacturing assembly actions on the HA4M dataset. Figure 8 represents the recognition summary of (a) convLSTM2D (b) LSTM with VG16 (c) LRCN (d) 3D CNN (e) ConvLSTM2D + 3D CNN (f) ConvLSTM2D + 3D CNN (Simplified) and figure 9 summarize the performance of the proposed hybrid model. The test evaluation is shown in Table 2 to measure the precision, recall, F1 score and test accuracy. Various approaches for action recognition were evaluated, and their performance metrics were assessed. The results indicate that 3D CNN + ConvLSTM2D achieved the highest precision (0.99445) and recall (0.99423), resulting in an impressive F1 Score of 0.99423. ConvLSTM2D and 3D CNN also demonstrated strong performance, while LSTM with VG16 had comparatively lower scores. These metrics offer valuable insights into the effectiveness of different methods for recognizing assembly actions, with the F1 Score serving as a comprehensive measure of overall performance.

The results of this study show that hybrid models are better suited for recognizing manufacturing assembly actions than baseline models. The reason to perform better

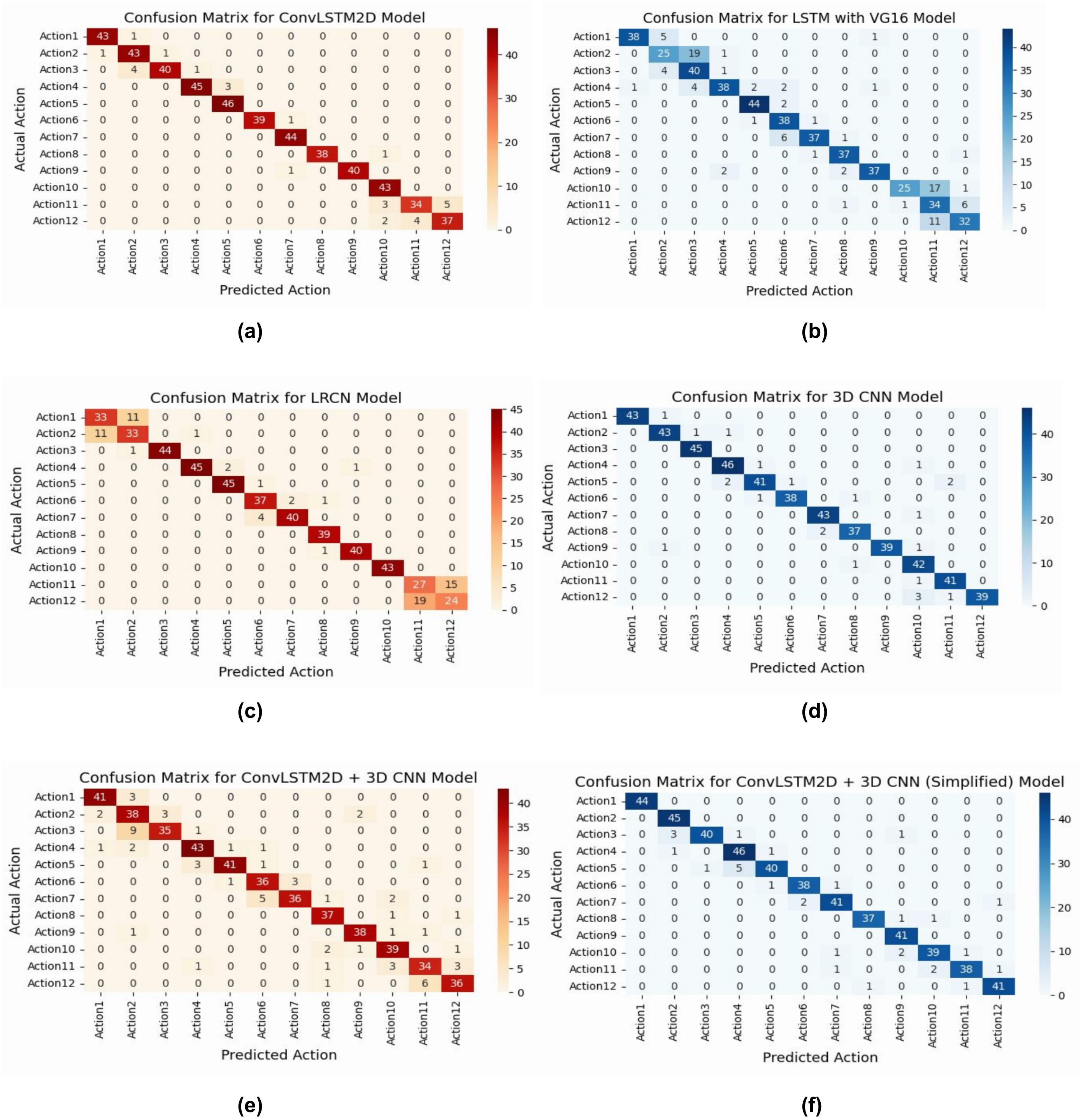

Figure 8: Confusion matrix of Assembly Action Recognition for: (a) convLSTM2D (b) LSTM with VG16 (c) LRCN (d) 3D CNN (e) ConvLSTM2D + 3D CNN (f) ConvLSTM2D + 3D CNN (Simplified)

Table 2: Evaluating Performance of Various Approaches for Assembly Action Recognition

Model	Precision	Recall	F1 Score	Test Accuracy (%)
ConvLSTM2D	0.94738	0.94615	0.94579	94.61
LSTM with VG16	0.84097	0.817731	0.81868	81.73
LRCN	0.86606	0.86538	0.86532	86.53
3D CNN	0.95161	0.95577	0.95594	95.57
ConvLSTM2D + 3D CNN	0.8774	0.87308	0.8736	87.30
ConvLSTM2D + 3D CNN (Simplified)	0.94357	0.94231	0.94208	94.23
3D CNN + ConvLSTM2D	**0.99445**	**0.99423**	**0.99423**	**99.42**

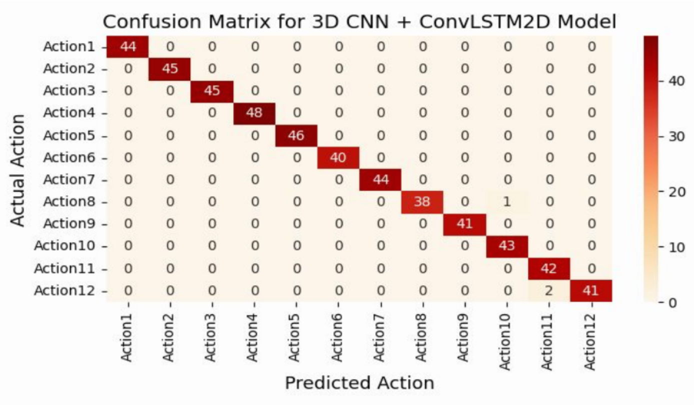

Figure 9: Confusion matrix of Assembly Action Recognition for proposed hybrid model (3D CNN + ConvLSTM2D)

is that hybrid models can better capture the temporal dependencies between video frames. The 3D CNN + ConvLSTM2D model performed the best because it combines the strengths of two different types of neural networks: 3D CNNs and ConvLSTM2Ds. 3D CNNs are good at extracting spatial features from video frames, while ConvLSTM2Ds extract temporal relationships from special features. Combining these two types of neural networks, the 3D CNN + ConvLSTM2D model can better capture the full range of features relevant to recognizing manufacturing assembly actions. The simplified version of the ConvLSTM2D + 3D CNN model performed well compared to its first version. It is possible to simplify the hybrid models without significantly impacting their performance. This is important because it can make the models more computationally efficient and easier to train. The results of this study provide a valuable starting point for future research on manufacturing assembly action recognition.

6. Conclusion

The growing importance of deep learning algorithms in human activity recognition has driven progress in intelligent manufacturing. The integration of artificial intelligence in manufacturing has gained attention in academic and industrial circles. Accurately recognizing assembly actions in smart manufacturing is essential for real-time efficiency assessment and monitoring. Despite limited research, recognizing human actions remains crucial to achieving Industry 4.0-aligned manufacturing goals. This study introduces a hybrid deep-learning model for monitoring assembly tasks and recognizing human actions in manufacturing. The model's performance is rigorously evaluated on the HA4M dataset, encompassing diverse actions in constructing an Epicyclic Gear Train. The research is organized into three domains: evaluating baseline models, developing a hybrid model combining convLSTM and 3D CNN, and introducing an efficient architecture using both networks. The outcome is an impressive 99.42% accuracy, highlighting the hybrid approach's potential for advancing understanding in intelligent manufacturing.

One drawback of utilizing RGB data to identify assembly actions is the substantial computational time required to train and test deep-learning models. Prospective research avenues encompass the following directions: (i) Hybrid models could be applied to alternative relevant datasets like IKEA ASM Dataset [24] and Assembly101 [25], etc., enabling the evaluation and comparison of model performance and efficacy. (ii) Developing a comprehensive multimodal deep learning framework could yield a robust assembly action recognition system capable of accommodating diverse data sources. (iii) Incorporating attention mechanisms and self-learning AI into assembly action recognition can potentially enhance future intelligent manufacturing endeavors. This enhancement could increase productivity, efficiency, sustainability of manufacturing, and human-robot collaboration in manufacturing.

ACKNOWLEDGMENT

We would like to acknowledge the financial support from the Bangabandhu Science and Technology Fellowship Trust, Bangladesh.

References

[1] G. Cicirelli, R. Marani, L. Romeo, M. G. Domínguez, J. Heras, A. G. Perri, T. D'Orazio, The ha4m dataset: Multi-modal monitoring of an assembly task for human action recognition in manufacturing, Scientific Data 9 (1) (2022) 745.

[2] T. Brogårdh, Present and future robot control development—an industrial perspective, Annual Reviews in Control 31 (1) (2007) 69–79.

[3] J. de Gea Fernández, D. Mronga, M. Günther, T. Knobloch, M. Wirkus, M. Schröer, M. Trampler, S. Stiene, E. Kirchner, V. Bargsten, et al., Multimodal sensor-based whole-body control

for human–robot collaboration in industrial settings, Robotics and Autonomous Systems 94 (2017) 102–119.

[4] G. Canal, S. Escalera, C. Angulo, A real-time human-robot interaction system based on gestures for assistive scenarios, Computer Vision and Image Understanding 149 (2016) 65–77.

[5] M. Al-Amin, R. Qin, M. Moniruzzaman, Z. Yin, W. Tao, M. C. Leu, An individualized system of skeletal data-based cnn classifiers for action recognition in manufacturing assembly, Journal of Intelligent Manufacturing (2021) 1–17.

[6] C. Chen, T. Wang, D. Li, J. Hong, Repetitive assembly action recognition based on object detection and pose estimation, Journal of Manufacturing Systems 55 (2020) 325–333.

[7] W. Tao, Z.-H. Lai, M. C. Leu, Z. Yin, Worker activity recognition in smart manufacturing using imu and semg signals with convolutional neural networks, Procedia Manufacturing 26 (2018) 1159–1166.

[8] M.-A. Zamora-Hernandez, J. A. Castro-Vargas, J. Azorin-Lopez, J. Garcia-Rodriguez, Deep learning-based visual control assistant for assembly in industry 4.0, Computers in Industry 131 (2021) 103485.

[9] A. Matin, M. R. Islam, X. Wang, H. Huo, G. Xu, Aiot for sustainable manufacturing: Overview, challenges, and opportunities, Internet of Things (2023) 100901.

[10] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[11] J. Zhang, P. Wang, R. X. Gao, Hybrid machine learning for human action recognition and prediction in assembly, Robotics and Computer-Integrated Manufacturing 72 (2021) 102184.

[12] T. Maekawa, D. Nakai, K. Ohara, Y. Namioka, Toward practical factory activity recognition: unsupervised understanding of repetitive assembly work in a factory, in: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2016, pp. 1088–1099.

[13] T. Stiefmeier, D. Roggen, G. Troster, Fusion of string-matched templates for continuous activity recognition, in: 2007 11th IEEE International Symposium on Wearable Computers, IEEE, 2007, pp. 41–44.

[14] M. Al-Amin, W. Tao, D. Doell, R. Lingard, Z. Yin, M. C. Leu, R. Qin, Action recognition in manufacturing assembly using multimodal sensor fusion, Procedia Manufacturing 39 (2019) 158–167.

[15] J. Wang, Y. Ma, L. Zhang, R. X. Gao, D. Wu, Deep learning for smart manufacturing: Methods and applications, Journal of manufacturing systems 48 (2018) 144–156.

[16] J. Berg, T. Reckordt, C. Richter, G. Reinhart, Action recognition in assembly for human-robot-cooperation using hidden markov models, Procedia CIRP 76 (2018) 205–210.

[17] D. Moutinho, L. F. Rocha, C. M. Costa, L. F. Teixeira, G. Veiga, Deep learning-based human action recognition to leverage context awareness in collaborative assembly, Robotics and Computer-Integrated Manufacturing 80 (2023) 102449.

[18] S. Li, J. Fan, P. Zheng, L. Wang, Transfer learning-enabled action recognition for human-robot collaborative assembly, Procedia CIRP 104 (2021) 1795–1800.

[19] H. Goto, J. Miura, J. Sugiyama, Human-robot collaborative assembly by on-line human action recognition based on an fsm task model, in: Human-robot interaction 2013 workshop on collaborative manipulation, 2013.

[20] K. Ikeuchi, T. Suchiro, Towards an assembly plan from observation. i. assembly task recognition using face-contact relations (polyhedral objects), in: Proceedings 1992 IEEE International Conference on Robotics and Automation, IEEE Computer Society, 1992, pp. 2171–2172.

[21] G. Z. Yang, L. Pei, K. Z. Xin, C. Z. Yi, Manual assembly action segmentation method based on spatiotemporal features, in: 2021 China Automation Congress (CAC), IEEE, 2021, pp. 6696–6701.

[22] E. Coronado, K. Fukuda, I. G. Ramirez-Alpizar, N. Yamanobe, G. Venture, K. Harada, Assembly action understanding from fine-grained hand motions, a multi-camera and deep learning approach, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021, pp. 2628–2634.

[23] F.-B. MARIN, G. GURĂU, M. MARIN, Real-time assembly operation recognition, The Annals of "Dunarea de Jos" University of Galati. Fascicle IX, Metallurgy and Materials Science 45 (4) (2022) 92–95.

[24] Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, S. Gould, The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 847–859.

[25] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, A. Yao, Assembly101: A large-scale multi-view video dataset for understanding procedural activities, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21096–21106.