Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 4916–4925

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Robust Multimodal Approach for Assembly Action Recognition

Abdul Matin[a], Md Rafiqul Islam[b], Xianzhi Wang[a], Huan Huo[a]

[a]*School of Computer Science, University of Technology Sydney, Australia*
[b]*Business Information Systems, Australian Institute of Higher Education (AIH), Australia*

Abstract

Human action recognition has been explored in healthcare, sports, and entertainment, with a recent shift toward manufacturing settings for monitoring assembly tasks. Identifying assembly actions is crucial for improving human-robot collaboration and optimizing the assembly process. However, the complexity of assembly tasks poses challenges for action recognition methods, with single-modality methods struggling to capture the complex dynamics and context. We proposed the multimodal ConvLSTM-AssNet and C3D-AssNet methods, which use RGB, RGB-A, and depth data. The models are tested in single, double, and triple stream configurations, with attention mechanisms integrated to focus on relevant features. The proposed models are evaluated on the HA4M dataset. Attention-Guided C3D-AssNet is most accurate for single (RGB-A: 97.10%) and double streams (RGB-A + Depth: 98.84%), while ConvLSTM-AssNet performs best for triple streams (RGB + RGB-A + Depth: 97.30%). This research advances multimodal assembly action recognition for manufacturing applications.

Peer-review under responsibility of the scientific committee of the 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems

Keywords:
Human Action Recognition; Deep Learning; Multimodal Assembly Action; Attention Mechanism; Smart Manufacturing

1. Introduction

Assembly action recognition is crucial for smart manufacturing systems to identify and categorize movements during assembly tasks. This technology can promote human-robot collaboration, optimize assembly processes, and boost productivity. Accurate identification of assembly actions can help manufacturers identify bottlenecks, detect inefficiencies, adjust production processes, allocate resources, and facilitate teamwork between human operators and robotic assistants.

* Corresponding author. Tel.: +61290208050
E-mail address: r.islam@aih.edu.au

1877-0509
Peer-review under responsibility of the scientific committee of the 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems
10.1016/j.procs.2024.09.448

Despite the advantages of traditional methods, using single-modality faces challenges in dealing with object interactions and detailed motion of actions [1], limiting their use in industrial settings. Assembly tasks often involve manipulations of parts, precise movements, and interactions with various tools and objects. Occlusions, varying lighting conditions, and cluttered backgrounds make it even harder for traditional recognition methods to be effective.

Multimodal deep learning methods can potentially capture the complicated dynamics of assembly actions and pull out important features [2]. Integrating contextual details from different modalities leads to more robust and precise recognition outcomes, addressing issues that single modalities have.

This study explores multimodal assembly action recognition. We designed ConvLSTM-AssNet and C3D-AssNet architectures, inspired by Convolutional Long Short Term Memory (ConvLSTM) [3] and 3D Convolutional Neural Networks [4]. The proposed AssNet approaches can merge essential features from RGB, RGB-A and depth modalities. The AssNet architectures are extended to double-stream and triple-stream variations for integrating modalities. We incorporate attention mechanisms to enable the networks to concentrate on the spatiotemporal features and avoid irrelevant data while highlighting essential features. The performance is evaluated on the HA4M dataset [5]. Our experiments show that multimodal approaches achieve cutting-edge results compared to single-modality baselines.

The key contributions include:

- Introducing multimodal assembly networks (AssNet), such as ConvLSTM-AssNet and C3D-AssNet, designed to handle multimodal data for assembly action recognition.
- Implementing attention mechanisms into our AssNets to extract the most essential features from the spatiotemporal regions.

The rest of the paper is organized as follows: Section 2 provides an overview of the related works; Section 3 explains the proposed architectures and attention mechanisms; Section 4 outlines the experimental setup and assessment criteria; Section 5 describes the results and comparison; finally, Section 6 wraps up the chapter while suggesting future avenues.

2. Literature Review

The manufacturing industry is transforming due to the adoption of Industry 4.0 technologies such as automation, the Internet of Things (IoT), and Artificial Intelligence (AI) [6]. This shift is leading to improvements in manufacturing assembly processes and enhancing product quality. Human Activity Recognition (HAR) has become an important component of eco-friendly manufacturing practices. HAR system can monitor and understand workers' activities in manufacturing settings by capturing assembly actions, movements, and interactions [7]. It can enable real-time assembly process optimization, quality control and safety measures [8]. HAR system can effectively boost productivity, efficiency and sustainability by offering insights into process bottlenecks, inefficiencies and areas that can be enhanced [9]. Ensuring workers' safety and well-being in manufacturing environments is also crucial. HAR systems can monitor workers' motions and postures using sensors and AI. These systems offer feedback and alerts to prevent accidents and promote practices in the workplace [10]. It also facilitates collaboration between humans and machines. By understanding actions and anticipating human movements, HAR enables communication and coordination between humans and collaborative robots (cobots) naturally and intuitively [11].

However, incorporating HAR into manufacturing poses challenges due to its complex nature, various activities, tools, materials, machinery, noise levels, vibrations, and changing lighting conditions [8]. Accurate classification and object detection manipulate assembly actions, and even motions, and high-resolution sensors and lightweight algorithms are needed to capture movements with precision [12]. Nowadays, researchers are trying to investigate these difficulties. They have been implementing deep learning-based approaches to extract complex spatiotemporal features.

Deep learning has revolutionized the field of human activity recognition (HAR) by enabling the automatic learning of hierarchical and discriminative features from raw input data. Especially some deep learning methods, such as CNNs and RNNs, have shown impressive abilities to understand human activity [13]. CNNs have been applied to various types of sensor data, including images, videos, and time series signals [14]. For video-based HAR, 3D CNNs have been proposed to capture spatial and temporal features simultaneously [15]. Chen et al. [16] proposed a two-stage

network that could use features from different visual modalities and showed changes over time to find and predict small-scale assembly tasks in smart factories. The authors demonstrated that their method worked well on a dataset of industrial assembly tasks; however, model evaluation is limited to small-scale datasets.

According to Moutinho et al. [15], they proposed a ResNet-34 and LSTM-based approach to recognize human actions and get the big picture of an industrial engine assembly process. They used RGB-D and skeleton data to train and test the model. They demonstrated good performance; however, the model was not generalized to other operators in the training data and lacked variable datasets with multiple operators. LSTMs have been explored to recognize assembly actions [17]. For instance, Wang et al. [18] proposed ResNet and LSTM-based approach for recognizing assembly actions from multimodal sensor data. The authors showed the significance of their method in capturing temporal dependencies and achieving high recognition accuracy on an industrial assembly dataset.

In HAR, multimodal fusion deployment extracts the essential features and merges features from sources for precise predictions. There are three fusion strategies: early fusion (feature level), late fusion (decision level), and intermediate fusion (hybrid). Early fusion combines features from different modalities before putting them into a classification or regression model [19]. Late fusion involves training models for each modality and merging their predictions through averaging voting or weighted sum. On the other hand, intermediate fusion merges the benefits of late fusion by integrating at various stages, enabling cross-modal interaction learning while retaining modality-specific details [20].

Numerous fusion strategies have been used for HAR in manufacturing. Al-Amin et al. [21] suggested a multimodal approach that uses sensor fusion algorithms to improve the recognition of human actions in factory assembly tasks. This approach combines data from EMG, IMU, and Kinect sensors. CNN models were trained on data from each type of sensor, and their outputs were combined using several fusion methods. The weighted fusion method worked best for identifying specific actions. However, the model's performance is limited to a specific assembly process and lacks scalability for various workstations. Wang et al. [22] created a new cross-domain few-shot learning method for detecting multimodal human actions in situations where humans and robots work together to put things together. They used a hierarchical data fusion mechanism to merge skeletal, RGB image, and depth map data modalities.

Assembly action recognition through applying attention mechanisms has emerged as a promising approach. When attention mechanisms are added to deep learning models, they can effectively find and group the steps taken during the assembly process by focusing on the most important spatial and temporal features of input data, such as video frames or sensor readings. Fine-grained action segmentation is critical for enhancing work efficiency and identifying worker errors in assembly scene analysis. However, the intricate nature of the actions involved poses significant challenges. Chen et al. [23] proposed a model with an attention mechanism for focusing on important and useful parts and extracting essential features. By selectively attending to key visual cues, the attention mechanism enhances the model's ability to capture and interpret the subtle nuances and intricate details essential for accurately recognizing assembly behaviors. A study by [24] proposed a STAR network that would use an attention mechanism to group video clips together based on class labels, removing noise from background or other actions, and an enhanced recurrent neural network to figure out how actions happen in relation to each other in time.

In summary, this study emphasizes the significance of HAR in manufacturing environments to improve assembly processes and collaboration between humans and robots. It also underscores the value of deep learning methods and multimodal fusion strategies for recognizing assembly activities. Additionally, this study recognizes the complexity involved in assembly actions within manufacturing setups.

3. Methodologies

This section outlines multimodal approaches for assembly action recognition across multimodal data using advanced deep-learning models. We introduce two architectures, ConvLSTM-AssNet and C3D-AssNet, to capture spatiotemporal relationships and integrate data from various sources such as RGB, RGB-A (alpha channel), and depth data.

3.1. ConvLSTM-AssNet Architecture

The basic structure of ConvLSTM-AssNet is shown in Figure 1. It consists of four ConvLSTM blocks, and each one has a ConvLSTM2D layer, MaxPooling3D operations, and TimeDistributed Dropout layers for robust feature

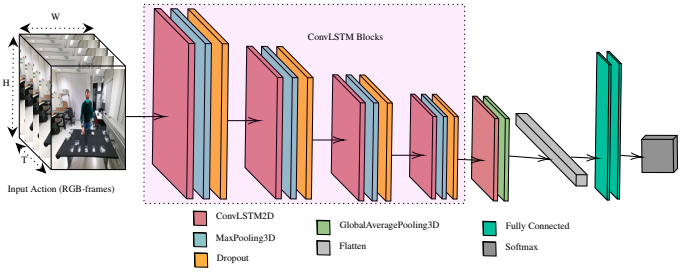

Fig. 1. ConvLSTM-AssNet base model architecture: Input action consists of input shape (40x64x64x3), where 40 represents T sequence length, 64 represents spatial dimensions H and W, and 3 represents RGB channels.

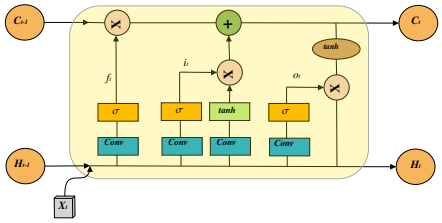

Fig. 2. Architecture of ConvLSTM2D cell

extraction and understanding of how temporal dependencies change over time. A single ConvLSTM2D layer with GlobalAveragePooling3D reduces spatio-temporal information, followed by a Flatten layer and fully connected layers for assembly action recognition. The ConvLSTM2D layers use a 3x3 kernel size, tanh activation, a recurrent dropout of 0.2, and increasing filters from 4 to 32.

ConvLSTM2D extends LSTM to handle spatial-temporal data through convolutional operations while capturing temporal associations. The fundamental architecture of the ConvLSTM2D cell is shown in Figure 2. Given an input tensor $\mathbf{X}_t \in \mathbb{R}^{m \times n \times c}$ at time step t, with spatial dimensions m and n, and c channels, ConvLSTM2D maintains a state $\mathbf{H}_t \in \mathbb{R}^{m \times n \times h}$ and a cell state $\mathbf{C}_t \in \mathbb{R}^{m \times n \times h}$, where h is the number of filters. The layer updates these states through operations:

$$i_t = \sigma(\mathbf{W}_{xi} * \mathbf{X}_t + \mathbf{W}_{hi} * \mathbf{H}_{t-1} + \mathbf{b}_i) \tag{1}$$

$$f_t = \sigma(\mathbf{W}_{xf} * \mathbf{X}_t + \mathbf{W}_{hf} * \mathbf{H}_{t-1} + \mathbf{b}_f) \tag{2}$$

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_{xc} * \mathbf{X}_t + \mathbf{W}_{hc} * \mathbf{H}_{t-1} + \mathbf{b}_c) \tag{3}$$

$$\mathbf{C}_t = f_t \odot \mathbf{C}_{t-1} + i_t \odot \tilde{\mathbf{C}}_t \tag{4}$$

$$o_t = \sigma(\mathbf{W}_{xo} * \mathbf{X}_t + \mathbf{W}_{ho} * \mathbf{H}_{t-1} + \mathbf{b}_o) \tag{5}$$

$$\mathbf{H}_t = o_t \odot \tanh(\mathbf{C}_t) \tag{6}$$

Where $*$ is convolution, $\odot$ is element-wise multiplication, $\sigma(\cdot)$ and $\tanh(\cdot)$ are activation functions, and $\mathbf{W}_{x*} \in \mathbb{R}^{k \times k \times c \times h}$, $\mathbf{W}_{h*} \in \mathbb{R}^{k \times k \times h \times h}$, and $\mathbf{b}_* \in \mathbb{R}^h$ are learning parameters.

The output goes through MaxPooling3D (pool size 1x2x2), TimeDistributed Dropout, another ConvLSTM2D layer with 64 filters, GlobalAveragePooling3D, Flatten, and Dense layers with Softmax activation for class probabilities. ConvLSTM-AssNet is extended to double-stream and triple-stream architectures shown in Figures 3 to extract features from RGB, RGB-A, and depth data in parallel streams before fusing for assembly action recognition.

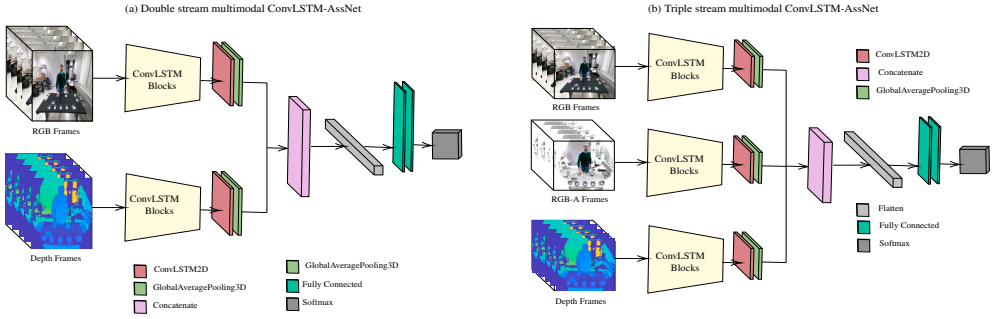

Fig. 3. Double and Triple-stream multimodal ConvLSTM-AssNet Architectures

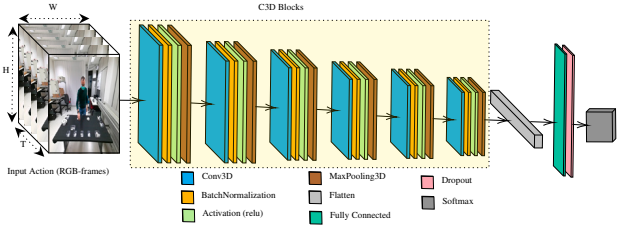

Fig. 4. C3D-AssNet base model architecture

3.2. C3D-AssNet Architecture

The C3D-AssNet methodology utilizes 3D Convolutional Neural Networks (C3D) [4] to learn discriminative spatiotemporal features from raw input frame sequences. The base model shown in Figure 4 consists of six C3D blocks, each containing convolutional layers, batch normalization, activation functions, and max pooling layers. The input is an assembly action of shape (T, H, W, C) denoting (40, 64, 64, 3), where T is the number of frames, H and W are spatial dimensions, and C is the number of channels.

For an input action sequence $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$, the first convolutional layer applies 3D convolution with learnable filters $\mathbf{W}_1 \in \mathbb{R}^{t\times d\times d\times C\times f_1}$, where t is temporal dimension, d is spatial dimension, and f_1 is the number of filters, progressively increasing from 16 to 512 across layers. The output is denoted as $\mathbf{X}_1$. Subsequent layers increase filters and apply batch normalization, ReLU activation, and Maxpooling with pool size (1, 2, 2) for the first block and (2, 2, 2) for the rest, and strides (1, 2, 2) for the first block and (2, 2, 2) for the rest, producing output $\mathbf{X}_p$. A fully connected layer with output $\mathbf{X}_{\text{fc}}$ is followed by a Softmax layer computing predicted probabilities $\hat{\mathbf{y}}$:

$$\mathbf{X}_1 = \text{Conv3D}(\mathbf{X}, \mathbf{W}_1) + \mathbf{b}_1 \tag{7}$$

$$\mathbf{X}_p = \text{MaxPool3D}(\mathbf{X}_l) \tag{8}$$

$$\mathbf{X}_{\text{fc}} = \text{ReLU}(\mathbf{W}_{\text{fc}} \cdot \mathbf{X}_{\text{flat}} + \mathbf{b}_{\text{fc}}) \tag{9}$$

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_{\text{out}} \cdot \mathbf{X}_{\text{fc}} + \mathbf{b}_{\text{out}}) \tag{10}$$

Similar to ConvLSTM-AssNet, C3D-AssNet is extended to double-stream and triple-stream architectures (Figure 5) by processing multiple modalities in parallel streams and fusing their outputs for assembly action recognition.

3.3. Attention-Guided Feature Learning

To enhance the performance of our model, we incorporated an attention mechanism that draws inspiration from the Convolutional Block Attention Module (CBAM) [25]. This feature blends spatiotemporal attention used to enhance the feature representation and extract significant features and channels. Given an input tensor $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$, the

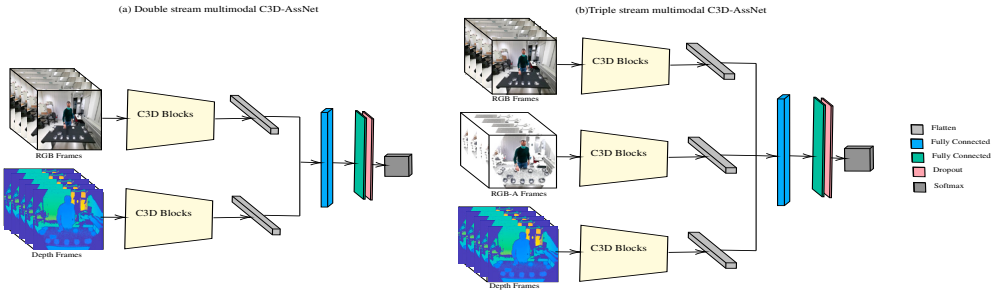

Fig. 5. Double and Triple stream multimodal C3D-AssNet architectures

channel attention computes average-pooled ($\mathbf{F}_{\text{avg}}$) and max-pooled ($\mathbf{F}_{\text{max}}$) features across spatial dimensions:

$$\mathbf{F}_{\text{avg}} = \frac{1}{T \times H \times W} \sum_{i=1}^{T} \sum_{j=1}^{H} \sum_{k=1}^{W} \mathbf{X}_{i,j,k} \tag{11}$$

$$\mathbf{F}_{\text{max}} = \max_{i,j,k}(\mathbf{X}_{i,j,k}) \tag{12}$$

These features are passed through convolutional layers with a reduction ratio r to obtain channel attention weights $\mathbf{M}_c$, which are multiplied with the input tensor $\mathbf{X}$:

$$\mathbf{F}'_{\text{avg}} = \text{ReLU}(\text{Conv3D}_{1\times1\times1}(\mathbf{F}_{\text{avg}}, \frac{C}{r})) \tag{13}$$

$$\mathbf{F}'_{\text{max}} = \text{ReLU}(\text{Conv3D}_{1\times1\times1}(\mathbf{F}_{\text{max}}, \frac{C}{r})) \tag{14}$$

$$\mathbf{M}_c = \sigma(\text{Conv3D}_{1\times1\times1}(\mathbf{F}'_{\text{avg}}, C) + \text{Conv3D}_{1\times1\times1}(\mathbf{F}'_{\text{max}}, C)) \tag{15}$$

$$\mathbf{X}_c = \mathbf{X} \odot \mathbf{M}_c \tag{16}$$

The spatial attention computes channel-wise average-pooled ($\mathbf{F}_{s,\text{avg}}$) and max-pooled ($\mathbf{F}_{s,\text{max}}$) features, concatenates them, and passes them through a convolutional layer with a $7 \times 7 \times 7$ kernel to obtain spatial attention weights $\mathbf{M}_s$:

$$\mathbf{F}_{s,\text{avg}} = \frac{1}{C} \sum_{i=1}^{C} \mathbf{X}_{c,:,:,:,i} \tag{17}$$

$$\mathbf{F}_{s,\text{max}} = \max_{i}(\mathbf{X}_{c,:,:,:,i}) \tag{18}$$

$$\mathbf{M}_s = \sigma(\text{Conv3D}_{7\times7\times7}([\mathbf{F}_{s,\text{avg}}; \mathbf{F}_{s,\text{max}}], 1)) \tag{19}$$

$$\mathbf{X}_{\text{cs}} = \mathbf{X}_c \odot \mathbf{M}_s \tag{20}$$

The output $\mathbf{X}_{\text{cs}}$ is an attention-refined representation focusing on relevant channels and spatial locations.

We incorporated this attention mechanism into our ConvLSTM-AssNet and C3D-AssNet models. The attention-guided ConvLSTM-AssNet model (Figure 6) consists of three ConvLSTM blocks followed by an attention layer. The last ConvLSTM layer is replaced with an attention layer followed by AveragePooling3D. Similarly, in the C3D-AssNet model (Figure 4), we used four C3D blocks, and instead of the last two C3D blocks, we used an attention

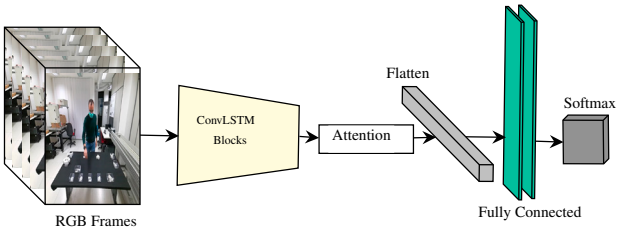

Fig. 6. Attention-Guided ConvLSTM-AssNet

layer followed by an AveragePooling layer. The attention mechanism is consistently applied across the single, double, and triple-stream architectures, with separate attention layers for each modality stream in the multi-stream variants.

4. Experimental Setup

4.1. Dataset Preprocessing

We employed the HA4M [5] dataset for training and testing models. It contains multi-modal data capturing actions performed by 41 subjects during an Epicyclic Gear Train (EGT) assembly in a laboratory setting using a Microsoft Azure Kinect. The dataset includes 12 actions performed by each individual across many trials and consists of six data types: RGB frames, depth maps, IR frames, RGB-to-Depth-Aligned frames, point clouds, and skeleton data. Our experiments utilized RGB, RGB-to-Depth-Aligned (RGB-A), and depth map data modes. Data preprocessing involved capturing frames from assembly actions, adjusting their size, standardizing them, and arranging them into an organized dataset with a sequence length of T = 40. If the number of frames exceeded this length, a frame-skipping mechanism was used to distribute the frame selection evenly. Each selected frame was resized, normalized, and appended to a list. Zero padding was used if the number of frames was less than the sequence length. The resulting dataset comprised features, labels and action file paths, enabling supervised learning for assembly action recognition models.

4.2. Data Split for Experiments

We divided the preprocessede dataset into three segments to train and evaluate the AssNet models: training (80%), validation (10%), and testing (10%). The training set helps the models learn distinguishable features. The validation set serves as a training checkpoint, helping to alter parameters and prevent overfitting. This split ratio is widely used for deep learning-based action classification. Maintaining a class distribution across subsets is essential to avoid class imbalances and ensure proper evaluation.

4.3. Model Hyperparameter

We tuned hyperparameters for the ConvLSTM-AssNet and C3D-AssNet models based on how well they worked and what we knew about the assembly action recognition domain. The hyperparameter tuning includes learning rate, batch size, epochs, optimizer, dropout rates, activation functions, early stopping criteria, and batch normalization. The tuning process included trying out approaches and checking them against a validation set to determine the setup that would improve test results. Although the values used were specific, in our tests, there is room for refinement and exploration of advanced methods to pinpoint the most efficient configurations.

4.4. Model Architecture Variants

We explored single-stream base models, double-stream and triple-stream multimodal architectures and attention-guided variants of ConvLSTM-AssNet and C3D-AssNet. We tested these versions for the best multimodal AssNet performance and attention mechanisms for assembly action recognition.

Table 1. Single Stream Performance

Backbone	Model input	Precision	Recall	F1-score	Accuracy
LRCN	RGB	0.8802	0.8086	0.8121	0.8069
	RGB-A	0.905	0.8729	0.863	0.8764
	Depth	0.7463	0.6508	0.6895	0.6564
I3D	RGB	0.5811	0.4944	0.5115	0.4942
	RGB-A	0.6438	0.5039	0.5531	0.5058
	Depth	0.638	0.5029	0.5878	0.5012
ConvLSTM-AssNet	RGB	0.8853	0.8765	0.8781	0.8803
	RGB-A	0.9541	0.9468	0.9476	0.9498
	Depth	0.9292	0.9256	0.9251	0.9266
C3D-AssNet	RGB	0.941	0.9202	0.9174	0.9228
	RGB-A	0.9507	0.9512	0.9506	0.9507
	Depth	0.9069	0.858	0.8754	0.8533

4.5. Evaluation Matrix

We used precision, recall, F1-score, and accuracy metrics to evaluate model performance. Precision measures the ability to classify positive instances accurately; recall represents the capacity to detect positive occurrences; the F1-score balances precision and recall, and accuracy provides an overall measure of correctness. Integrating these metrics offers a comprehensive view of model performance.

5. Result and Discussion

The experimental results in Tables 1, 2, 3, and 4 evaluate the performance of different models and input combinations for assembly action recognition using precision, recall, F1-score, and accuracy metrics.

5.1. Single Stream Performance

In the single-stream setting (Table 1), ConvLSTM-AssNet and C3D-AssNet outperformed LRCN and I3D, highlighting their superior ability to capture spatiotemporal dependencies and extract discriminative features. Attention-guided variants further improved performance, suggesting the effectiveness of attention mechanisms in focusing on informative features. For C3D-AssNet, RGB and RGB-A modalities yielded the highest results, while ConvLSTM-AssNet performed best with depth data.

5.2. Double Stream Performance

Transitioning to the double-stream setting (Tables 2 and 4), fusing RGB and RGB-A modalities consistently achieved the highest scores across all models, indicating their complementary nature. Attention-guided models, particularly Attention-Guided C3D-AssNet, further enhance performance by selectively focusing on informative aspects of the input data.

5.3. Triple Stream Performance

Incorporating RGB, RGB-A, and Depth modalities (Tables 3 and 4), ConvLSTM-AssNet achieved remarkable accuracy of 97.3%, outperforming its counterparts. However, Attention-Guided ConvLSTM-AssNet exhibited a performance decline compared to its dual-stream variant, suggesting compromised effectiveness with an additional modality. Conversely, Attention-Guided C3D-AssNet emerged as the top-performing model.

5.4. Performance Comparison

A clear progression of performance improvement was observed as models transitioned from single-stream to double-stream and triple-stream settings, highlighting the significance of integrating multiple modalities for capturing meaningful features. Attention-guided variants consistently outperformed their non-attention counterparts across all settings, except for Attention-Guided ConvLSTM-AssNet in the triple-stream configuration.

Table 2. Double Stream Performance

Backbone	Model input	Precision	Recall	F1-score	Accuracy
LRCN	RGB + RGB-A	0.923	0.9133	0.9123	0.9151
	RGB + Depth	0.908	0.8524	0.8467	0.8571
	RGB-A + Depth	0.889	0.8688	0.862	0.8687
I3D	RGB + RGB-A	0.724	0.6298	0.6656	0.6293
	RGB + Depth	0.7146	0.6382	0.6588	0.6409
	RGB-A + Depth	0.6934	0.5993	0.6108	0.5946
ConvLSTM-AssNet	RGB + RGB-A	0.9343	0.9241	0.9244	0.9266
	RGB + Depth	0.95	0.9459	0.9463	0.9459
	RGB-A + Depth	0.9361	0.928	0.9289	0.9305
C3D-AssNet	RGB + RGB-A	0.9672	0.9626	0.9634	0.9614
	RGB + Depth	0.9401	0.9327	0.9345	0.9344
	RGB-A + Depth	0.9823	0.9807	0.981	0.9807

Table 3. Triple Stream Performance

Backbone	Model input	Precision	Recall	F1-score	Accuracy
LRCN	RGB + RGB-A + Depth	0.9241	0.9026	0.9045	0.9035
I3D		0.7075	0.6341	0.6608	0.6371
ConvLSTM-AssNet		0.9744	0.9745	0.974	0.973
C3D-AssNet		0.9684	0.965	0.9658	0.9653

Table 4. Attention-Guided ConvLSTM-AssNet and C3D-AssNet Performance

Model Architecture		Attention-Guided ConvLSTM-AssNet				Attention-Guided C3D-AssNet			
Model Name	Model input	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Single Stream	RGB	0.9273	0.915	0.9198	0.917	0.9714	0.9713	0.971	0.971
	RGB-A	0.955	0.9536	0.9539	0.9537	0.9749	0.9697	0.9715	0.971
	Depth	0.9702	0.9713	0.9705	0.971	0.9262	0.9164	0.9158	0.9112
Double Stream	RGB + RGB-A	0.9728	0.9732	0.9727	0.973	0.9886	0.9882	0.9883	0.9884
	RGB + Depth	0.9656	0.9648	0.9647	0.9653	0.9863	0.9849	0.9855	0.9846
	RGB-A + Depth	0.9436	0.9422	0.9421	0.9421	0.989	0.9879	0.9884	0.9884
Triple Stream	RGB + RGB-A + Depth	0.9294	0.9229	0.9255	0.9247	0.9835	0.9814	0.9822	0.9826

The superior performance of multimodal AssNet architectures can be attributed to:

- Attention mechanisms focus on discriminative features while suppressing irrelevant information.
- C3D-AssNet captures spatiotemporal dependencies and learns robust features.
- Fusion of multiple modalities provides a comprehensive representation by extracting complementary features from RGB, RGB-A, and Depth streams.

Attention-guided C3D-AssNet, combining double and triple modalities, achieved state-of-the-art performance, showcasing its potential for real-world industrial applications. The insights gained pave the way for further research and development of advanced methods for assembly action recognition, facilitating automation and optimization of industrial processes.

6. Conclusion

This study presented the ConvLSTM-AssNet and C3D-AssNet structures. These architectures capture spatial and temporal data to recognize assembly actions in manufacturing settings. The multimodal combinations, such as double-stream and triple-stream models, were examined to determine the optimal RGB, RGB-A, and depth modalities for enhanced recognition performance. Integrating attention processes led to developing attention-guided ConvLSTM-AssNet and C3D-AssNet architectures, which improved the AssNet models. The test results on the HA4M dataset demonstrated that multimodal topologies outperformed single-modal baselines. The Attention-Guided C3D-AssNet demonstrated outstanding performance when applied to double and triple modalities. This study establishes the foundation for future research in assembly action recognition using multimodal data. Future research could look into semisupervised learning, sensor-fusion methods to combine multimodal data, models that can be explained and interpreted, and making real-time manufacturing systems more scalable.

Acknowledgements

This research was supported by the Bangabandhu Science and Technology Fellowship Trust (BSTFT), Ministry of Science and Technology, Bangladesh.

References

[1] F. Zhu, L. Shao, J. Xie, Y. Fang, From handcrafted to learned representations for human action recognition: A survey, Image and Vision Computing 55 (2016) 42–52.
[2] W. Tao, M. C. Leu, Z. Yin, Multi-modal recognition of worker activity for human-centered intelligent manufacturing, Engineering Applications of Artificial Intelligence 95 (2020) 103868.
[3] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems 28 (2015).
[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
[5] 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.
[6] A. Matin, M. R. Islam, X. Wang, H. Huo, G. Xu, Aiot for sustainable manufacturing: Overview, challenges, and opportunities, Internet of Things (2023) 100901.
[7] S. Knoch, N. Herbig, S. Ponpathirkoottam, F. Kosmalla, P. Staudt, D. Porta, P. Fettke, P. Loos, Sensor-based human–process interaction in discrete manufacturing, Journal on Data Semantics 9 (2020) 21–37.
[8] M. Dallel, V. Havard, Y. Dupuis, D. Baudry, Digital twin of an industrial workstation: A novel method of an auto-labeled data generator using virtual reality for human action recognition in the context of human–robot collaboration, Engineering applications of artificial intelligence 118 (2023) 105655.
[9] 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.
[10] N. D. Nath, T. Chaspari, A. H. Behzadan, Automated ergonomic risk monitoring using body-mounted sensors and machine learning, Advanced Engineering Informatics 38 (2018) 514–526.
[11] K.-J. Wang, C. J. Lin, A. A. Tadesse, B. H. Woldegiorgis, Modeling of human–robot collaboration for flexible assembly—a hidden semi-markov-based simulation approach, The International Journal of Advanced Manufacturing Technology 126 (11) (2023) 5371–5389.
[12] T. Wang, P. Zheng, S. Li, L. Wang, Multimodal human–robot interaction for human-centric smart manufacturing: A survey, Advanced Intelligent Systems 6 (3) (2024) 2300359.
[13] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, Y. Liu, Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities, ACM Computing Surveys (CSUR) 54 (4) (2021) 1–40.
[14] S. K. Challa, A. Kumar, V. B. Semwal, A multibranch cnn-bilstm model for human activity recognition using wearable sensor data, The Visual Computer 38 (12) (2022) 4095–4109.
[15] 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.
[16] H. Chen, N. Zendehdel, M. C. Leu, Z. Yin, Fine-grained activity classification in assembly based on multi-visual modalities, Journal of Intelligent Manufacturing (2023) 1–19.
[17] 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 (2023) 1–17.
[18] Z. Wang, J. Yan, Multi-sensor fusion based industrial action recognition method under the environment of intelligent manufacturing, Journal of Manufacturing Systems 74 (2024) 575–586.
[19] Z. Ahmad, N. Khan, Human action recognition using deep multilevel multimodal (M^2) fusion of depth and inertial sensors, IEEE Sensors Journal 20 (3) (2020) 1445–1455. doi:10.1109/JSEN.2019.2947446.
[20] T. Huynh-The, C.-H. Hua, N. A. Tu, D.-S. Kim, Physical activity recognition with statistical-deep fusion model using multiple sensory data for smart health, IEEE Internet of Things Journal 8 (3) (2021) 1533–1543. doi:10.1109/JIOT.2020.3013272.
[21] 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.
[22] T. Wang, Z. Liu, L. Wang, M. Li, X. V. Wang, Data-efficient multimodal human action recognition for proactive human–robot collaborative assembly: A cross-domain few-shot learning approach, Robotics and Computer-Integrated Manufacturing 89 (2024) 102785.
[23] C. Chen, X. Zhao, J. Wang, D. Li, Y. Guan, J. Hong, Dynamic graph convolutional network for assembly behavior recognition based on attention mechanism and multi-scale feature fusion, Scientific Reports 12 (1) (2022) 7394.
[24] Y. Xu, C. Zhang, Z. Cheng, J. Xie, Y. Niu, S. Pu, F. Wu, Segregated temporal assembly recurrent networks for weakly supervised multiple action detection, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 9070–9078.
[25] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.