information

MDPI

Article

STAR-3D: A Holistic Approach for Human Activity Recognition in the Classroom Environment

Vijeta Sharma [1,2], Manjari Gupta [2], Ajai Kumar [1] and Deepti Mishra [3,*]

[1] Centre for Development of Advanced Computing (C-DAC), Pune 411008, India; vijetak@cdac.in (V.S.); ajai@cdac.in (A.K.)
[2] Computer Science, DST Center for Interdisciplinary Mathematical Sciences, Institute of Science, Banaras Hindu University, Varanasi 221005, India; manjari@bhu.ac.in
[3] Educational Technology Laboratory, Department of Computer Science (IDI), NTNU—Norwegian University of Science and Technology, 2815 Gjøvik, Norway
* Correspondence: deepti.mishra@ntnu.no

Abstract: The video camera is essential for reliable activity monitoring, and a robust analysis helps in efficient interpretation. The systematic assessment of classroom activity through videos can help understand engagement levels from the perspective of both students and teachers. This practice can also help in robot-assistive classroom monitoring in the context of human–robot interaction. Therefore, we propose a novel algorithm for student–teacher activity recognition using 3D CNN (STAR-3D). The experiment is carried out using India's indigenously developed supercomputer PARAM Shivay by the Centre for Development of Advanced Computing (C-DAC), Pune, India, under the National Supercomputing Mission (NSM), with a peak performance of 837 TeraFlops. The EduNet dataset (registered under the trademark of the DRSTA^{TM} dataset), a self-developed video dataset for classroom activities with 20 action classes, is used to train the model. Due to the unavailability of similar datasets containing both students' and teachers' actions, training, testing, and validation are only carried out on the EduNet dataset with 83.5% accuracy. To the best of our knowledge, this is the first attempt to develop an end-to-end algorithm that recognises both the students' and teachers' activities in the classroom environment, and it mainly focuses on school levels (K-12). In addition, a comparison with other approaches in the same domain shows our work's novelty. This novel algorithm will also influence the researcher in exploring research on the "Convergence of High-Performance Computing and Artificial Intelligence". We also present future research directions to integrate the STAR-3D algorithm with robots for classroom monitoring.

Keywords: human action recognition (HAR); classroom activity recognition; student–teacher action; deep learning; artificial intelligence; high-performance computing; human–robot interaction (HRI)

check for updates

Citation: Sharma, V.; Gupta, M.; Kumar, A.; Mishra, D. STAR-3D: A Holistic Approach for Human Activity Recognition in the Classroom Environment. *Information* **2024**, *15*, 179. https://doi.org/10.3390/info15040179

Academic Editor: Ognjen Arandjelović

Received: 5 February 2024
Revised: 8 March 2024
Accepted: 13 March 2024
Published: 25 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Engagement in an educational context refers to the level of involvement, interest, and interaction between students and teachers during the learning process [1]. Analysing engagement levels is crucial for understanding the effectiveness of teaching methods and identifying areas for improvement [2]. In addition, recommendations can be provided to improve teaching methods or enhance student participation based on analyses, highlighting the key aspects of an interactive and engaging classroom environment. The research suggests [2] that teachers' motivational behaviour significantly predicts student engagement, and educators should consider student engagement and teacher–student interaction. Previous findings revealed that teacher efficacy had a positive effect on students' classroom engagement, implying that efficacious teachers were more effective in engaging students in the classroom [3]. Artificial intelligence makes it possible to personalise learning according to the needs and assimilation capacities of each person. For example, this new technology can offer student exercises adapted to their learning level [4].

It is worth mentioning here that the ICT-based education system [5] is boosting students' learning and teacher's effectiveness. As a step ahead, artificial intelligence (AI)-enabled education is now the new era [6] for personalised learning. As per UNESCO, artificial intelligence (AI) has the potential to address some of the biggest challenges in education (https://www.unesco.org/en/digital-education/artificial-intelligence accessed on 30 November 2021); therefore, innovating teaching and learning practices through AI is urgently needed. Moreover, 'AI and education: Guidance for policy-makers' (UNESCO 2021) [7] was developed by UNESCO within the framework of the implementation of the Beijing Consensus, aimed at fostering AI-ready policy-makers in education. It aims to generate a shared understanding of the opportunities offered by AI for education, as well as its implications for the essential competencies required by the AI era.

Artificial intelligence (AI) offers numerous benefits and opportunities to understand the engagement level of students and teachers and transforms the way students learn, teachers teach, and educational institutions operate [4,8,9]. It can be useful in applications from personalised learning to adaptive learning platforms [10]. This kind of analysis could contribute to developing more effective teaching strategies and personalised learning experiences and improve student outcomes by leveraging the capabilities of AI in combination with human action recognition (HAR) techniques. There are several benefits of AI-based HAR, which help to understand the student–teacher engagement level in the classroom: for example, real-time feedback, adaptive teaching strategies based on student feedback, and trend analysis.

Presently, the barrier to understanding this engagement level is the evaluation method. With the continuous development of AI deep-learning (DL) algorithms and the promotion of human action recognition (HAR) application scenarios, research into HAR based on deep learning has become a key field in recent years. Nowadays, robots can also play a vital role in smart classrooms, where robots are equipped with AI techniques for classroom engagement level evaluation [11]. It serves as an advanced tool to enhance the educational environment by providing real-time insights into students' interactions and behaviours. However, in order to carry this out, advanced AI algorithms need to be designed for integration with robots. Another method is to install a live camera in the classroom that records classroom videos, but this would further require human intervention and a lengthy observation period for every video clip, in addition to requiring the researcher to judge various actions in the classroom, shallow teaching methods, students' interests, etc. Such a manual method cannot analyse the quality and efficiency of massive, recorded videos in the classroom. Therefore, there is a need to develop the action recognition of students and teachers using AI techniques in an educational context. Existing studies [12,13] only focus on either the student or the teacher. There was a study [14] carried out in a classroom environment; however, it was focused only on student behaviour recognition and not on action recognition.

To overcome this problem, the aim of the present study is to propose a novel AI deep-learning-based algorithm for the activity recognition of student–teacher interactions using 3D CNN (STAR-3D). This article is based on Chapter 6 of the author's PhD thesis [15].

In STAR-3D, a deep-learning-based human recognition method identifies the teachers' and students' activities from the video dataset. This method automatically identifies the actions that can be used for evaluation, feedback, monitoring, and analysis. We first found a scene and determined whether it belonged to the student or the teacher using the single-shot detection (SSD) method and then used 3D CNN to classify the action. Additionally, we used an advanced generative adversarial network (GAN) for data augmentation, which generated scenes from various angles. Our research work uses the power of India's HPC system—the PARAM Shivay supercomputer, under the national supercomputing mission (NSM) (https://nsmindia.in/node/155, accessed on 13 November 2021). The key features of the proposed STAR-3D algorithm are as follows:

1. STAR-3D is capable of automatically analysing the various actions as per the action categories in the EduNet dataset;

2. It functions as an intelligent classroom monitoring system. These analysed activities can be presented to the school management dashboard in a score based or any other preferred format;
3. To our knowledge, STAR-3D is the first deep-learning-based method to classify students' and teachers' activities in a classroom environment;
4. This model can also be integrated with the humanoid robot to assist in the classroom to monitor students' and teachers' engagement, such as monitoring active participation, interactive classroom behaviour analysis, collaboration assessment during group activity, etc.

The rest of the paper is structured as follows: Section 2 introduces related works; Section 3 provides an overview of the STAR-3D framework. Section 4 presents the experimental framework, Section 5 shows and discusses the results, and Section 6 concludes our work and outlines future research directions.

2. Related Study

In a related survey, a vision-based human action recognition system using a deep-learning technique is proposed by Chang et al. [16], which can recognise human actions by retrieving information from colour videos, optical flow videos, and depth videos from the camera. This core research of HAR is not focused on the classroom; rather, it is based on activities in an indoor environment. Another study [17] demonstrated instructor activity recognition using the spatiotemporal feature and feedforward learning method to classify the actions in eight categories based on a self-developed dataset. These categories include "Walking", "Pointing towards the board", "Pointing towards the screen", "Using a mobile phone", "Using a laptop", "Reading notes", "Sitting" and "Writing on the board".

Zhang and Ni [18] developed a 3D convolutional neural network based on human behaviour recognition to analyse and understand the actions of individuals and the interaction between multiple people in videos. This applies to research rooms or classroom scenarios to focus on students' learning behaviour. Li et al. [19] applied deep-learning techniques to recognise only students' actions in the classroom on a self-developed dataset with 15 actions specific to students in a classroom environment. Cheng et al. [20] identified the following activities using a deep convolutional generative adversarial network (GAN): writing, standing, sitting, raising a hand, playing with a smartphone, looking around, and climbing on the table. In another work, Gang et al. [14] presented their work for teachers' behaviour recognition in the classroom environment using 3D bilinear pooling for teacher behaviour recognition (3D BP-TBR) and validated the result on a self-built dataset. The actions included the following: bowing to students, pointing to the blackboard, writing on the blackboard, cleaning the blackboard, operating the interactive whiteboard, inviting students to answer questions, walking around the classroom, operating realia. Jisi and Yin [13] proposed a feature fusion network for student behaviour recognition, which was validated on UCF101 [21], HMDB51 [22], and on real student behaviour data in education. In an another study, authors [23] proposed a 3D attitude estimation algorithm using the RMPE (regional multi-person pose estimation) algorithm coupled with a deep neural network that combines human pose estimation and action recognition for basketball training. This algorithm was applied to a college sports basketball course to explore the influence of this teaching mode on classroom teaching effectiveness. An RFID-based approach, distinct from the vision-based approach for HAR, is proposed by Qiu et al. [24] for classroom action recognition. The system pastes a label to the right side of the desk and judges the students' learning state by recognising four movements: raising the left hand, raising the right hand, nodding off, and holding the book. It utilises a multichannel attentional graph convolutional neural network (ATGCN) to deeply learn the phase and signal strength of actions and conduct action recognition.

The existing daily action recognition, classroom action datasets [14,21,22], and the methods [13,14,17,20,25,26] of action recognition of the student and teacher do not automatically identify both the students' and the teachers' activities within the educational

context of actual teaching. Additionally, these methods are limited by a smaller number of action categories in the classroom environment. Therefore, we have developed the EduNet (DRSTA^{TM}) (https://edunet-drsta.com/, accessed on 22 October 2021) dataset and STAR-3D algorithm for the action recognition of students and teachers. In this paper, we proposed a novel algorithm based on 3D CNN, which is capable of recognising the real-time activities of students and teachers in the classroom environment. This algorithm is trained and validated on a self-developed EduNet dataset [27], which is specially developed for student and teacher activities in the classroom.

3. Proposed Methodology

This section provides an in-depth description of our proposed deep-learning-based activities recognition approach in the context of student and teacher. Figure 1 shows the architecture diagram of the proposed STAR-3D algorithm.

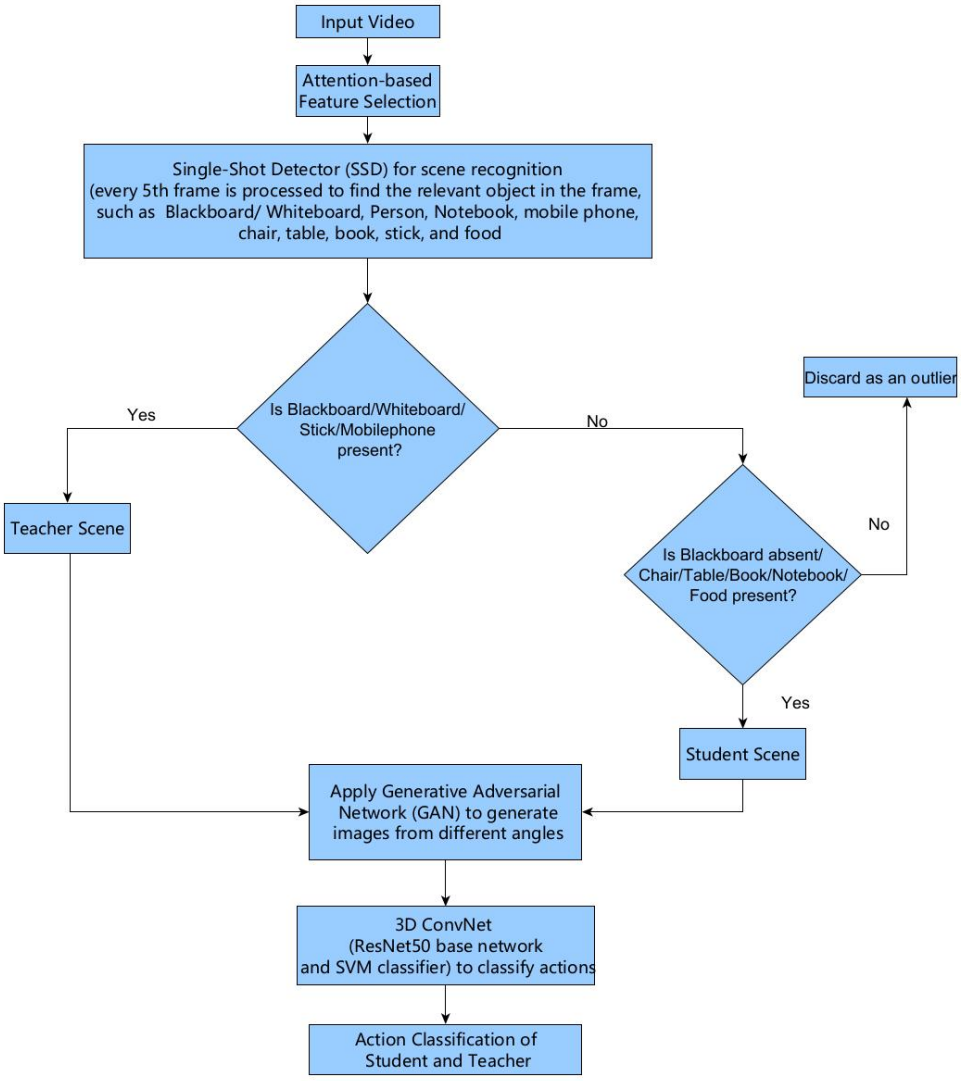


Figure 1. Architecture diagram of our proposed algorithm STAR-3D.

3.1. Attention-Based Student and Teacher Scene Recognition

An attention mechanism is employed to assign weights to each feature based on its relevance to the task. The attention mechanism learns to assign higher weights to more informative features and lower weights to less informative ones. These weights are then used to rank the features in terms of importance. Features with higher weights are considered more relevant and are selected to be part of the reduced feature subset. In STAR-

3D, the input video is processed using an attention-based feature-selection method [28], inspired by the concept of "attention" in human cognition, where certain elements are emphasised or focused on while others are ignored, i.e., automatic selection of relevant features from the image. It handles the blurred image generated by the video camera during the movement of people in the classroom. The presence of blur significantly reduces the image quality and then affects advanced visual task processing such as image segmentation, target recognition, and object detection. Afterwards, we applied a single-shot multibox detector (SSD) [29] to every fifth frame for scene recognition in the video. SSD is designed to efficiently detect and localise objects within images. The key idea behind SSD is to predict object bounding boxes and class labels directly from different feature maps extracted at various layers of a convolutional neural network (CNN). SSD is a single-pass algorithm that predicts object classes and bounding boxes directly, eliminating the need for multiple passes or region proposal networks. Known for its speed and efficiency, SSD is well-suited for real-time applications, especially for video analysis. We utilise SSD to identify the objects in the frame such as blackboards/whiteboards, people, notebooks, mobile phones, chairs, tables, books, sticks, and food. Figure 2 shows the architecture of the SSD object detector.

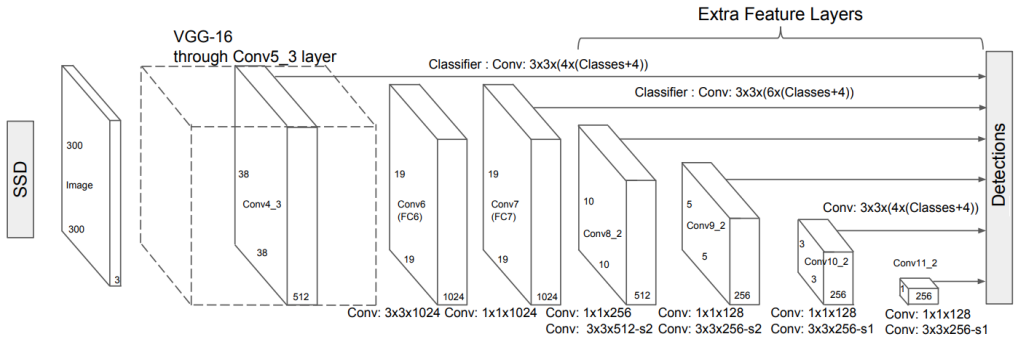


Figure 2. Architecture diagram of single-shot detector.

We trained the SSD model on our custom dataset, which consists of images extracted from videoclips in the EduNet dataset and these images were annotated with the Labelbox 3.13.0 (https://labelbox.com/, accessed on 16 November 2021) tool. A scene is categorised as a "teacher scene" if a blackboard, whiteboard, or stick is detected. If none of these items are detected, and there are no chair, table, book, notebook, or food items present, then the scene is categorised as a "student scene".

3.2. Data Generator

A generative adversarial network [30], or GAN, is a type of neural network architecture for generative modelling. A GAN is a generative model, trained with two neural network models. One model is the "generator" or "generative network" model which generates new plausible samples. Another is the "discriminator" or "discriminative network" model to distinguish between generated examples from real examples. Figure 3 shows the complete system of a GAN. The key idea behind GANs was to train a generator network to create data that are indistinguishable from real data, while simultaneously training a discriminator network to differentiate between real data and data generated by the generator. We used the GAN to generate more samples from various angles. This allowed us to create realistic classroom images resembling blackboards, sticks, desks, mobile phones, and more relevant images. We used it to generate additional training data for our tasks. Additionally, GAN helped to enhance the resolution of existing and generated images, turning low-resolution images into high-resolution counterparts. This approach enabled us to increase the volume of data while maintaining high resolution.

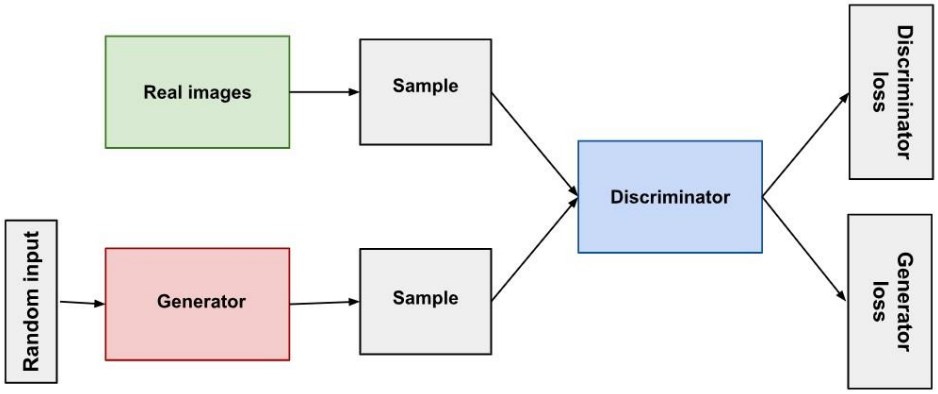


Figure 3. A whole system of a generative adversarial network (GAN).

3.3. Action Recognition of Student and Teacher 3D Network Architecture

STAR-3D is a two-step framework for human action recognition in a classroom environment. The first step focuses on scene detection in the classroom and categorises scenes as either a "student scene" or "teacher scene". For this, we utilised a single-shot multibox detector (SSD), as described in Section 3.1.

The second step is inspired by Tran et al. [31] and involves a 3D convolutional neural network to preserve the spatiotemporal feature of video data for action recognition. At the end of the network, we employed a support vector machine (SVM) as an action classifier. Figure 4 depicts the complete view of ResNet50 architecture with ReLu and batch normalisation. An important aspect of the framework is action data augmentation, for which we employed the generative adversarial network (GAN) [30], explained in Section 3.2. Additionally, we utilised the Kalman filter [32] to optimally estimate the variables of interest and Adam optimiser [33] to dynamically adjust the learning rates for each parameter during training. The neural network used for 3D CNN is a derivative of the ResNet50 network architecture. The last layer of this network is a fully connected layer: a combination of a single convolution layer, batch normalisation, max pooling, and a dropout layer. We sampled every independent convolutional layer with the activation function used to find dependencies between feature matrices, which returns the hidden state of neurons. We used leaky ReLu instead of simple ReLu as an activation function and a softmax fully connected layer. Given the large number of features to process, we utilised a mini-batch gradient descent (MBGD) to reduce the computation cost and speed up gradient computation. Additionally, mini-batch gradient descent can converge faster than batch gradient descent as it takes smaller steps towards the minimum of the cost function, helping the algorithm to escape from local minima.

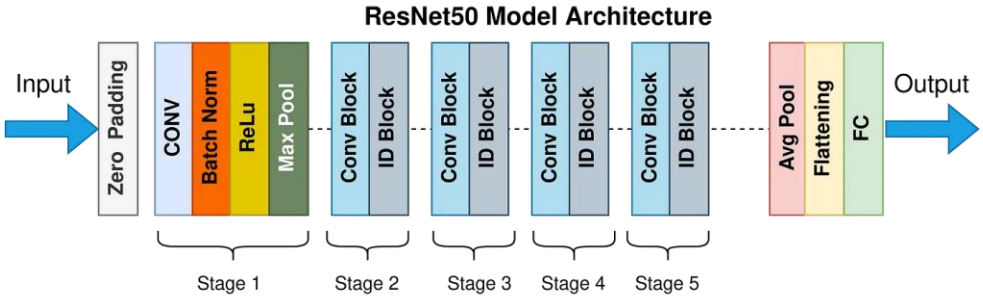


Figure 4. ResNet50 model architecture as the base model of STAR-3D.

Leaky ReLu: The fully connected layer with the activation function is used as a hidden layer to find a pattern between the hidden states of the previous layer neurons and the

output. We utilise the leaky ReLu function instead of the ReLu activation function to define the output of batch normalisation layers, as shown in Equation (1).

$$y = \{x,\ x \geq 0 \quad and \quad x * A,\ x < 0\} \tag{1}$$

Here, 'y' presents the output of the ReLu function and 'x' is the input value. 'A' is the constant defined in the function and usually has a very small value such as 0.01 or 0.05.

Next, we incorporated an attention mechanism to recognise the most relevant parts of action in the videos and to produce a salient discriminative representation of the HAR. Additionally, we compute the centre loss together with the softmax to produce the final probability and classify the actions collectively. The fused loss method is applied to improve the prediction accuracy. These loss functions are calculated after the fully connected network layer.

Softmax: The fully connected layer with the activation function used for assigning a video fragment to a certain class.

4. Experimental Framework

In this experiment, we utilised the EduNet dataset, which is developed by the authors and comprises 20 action categories deemed applicable for analysing the interactive classroom environment. The proposed AI model STAR-3D was implemented using a 3D convolutional neural network (CNN) architecture with a single-shot detector. Data augmentation techniques such as GAN, including random rotations and flips, were applied to enhance the model's robustness. The experiments were conducted on the India's HPC system PARAM Shivay (https://nsmindia.in/, accessed on 13 November 2021) with a total peak performance of 837 TFLOPS and equipped with NVIDIA Tesla V100 (designed and developed by NVIDIA, Santa Clara, CA, USA). We used 2 GPUs to train the model in a parallel manner. TensorFlow version 2.4 was used as a deep-learning framework.

4.1. Datasets

The STAR-3D algorithm has been trained and validated on a self-developed EduNet dataset [27]. This dataset consists of 20 action classes containing around 7851 manually annotated clips recorded in the actual classroom environment and extracted from YouTube videos. Each action category contains a minimum of 200 clips, totalling approximately 12 h of video data. EduNet is the first dataset specifically designed for monitoring activities in classrooms, including those of teachers and students. These video clips in the dataset were manually recorded in the 1st to 12th standard North Indian rural school classrooms, as well as extracted from YouTube videos. Annotations of each clip were also performed manually following the same annotation pattern as the UCF101 benchmark HAR dataset. Table 1 provides specifications of the EduNet dataset and Table 2 lists the 20 action categories included in this dataset.

Table 1. EduNet dataset specification.

Characteristic	Description
Actions	20
Clips	7851
Min clip length	3.25 s
Max clip length	12.7 s
Total duration	12 h
Minimum clip per class	200
Maximum clip per class	593
Frame rate	30
Resolution	1280 × 720
Audio	Yes
Coloured	Yes

Table 2. The 20 action classes of the EduNet dataset.

S.No.	Class	Action Belongs to Teacher/Student	Number of Clips
1	Arguing	Student	480
2	Clapping	Student	393
3	Eating_in_Classroom	Student	361
4	Explaining_the_Subject	Teacher	611
5	Gossip	Student	218
6	Hand_Raise	Student	389
7	Hitting	Teacher	387
8	Holding_Books	Teacher	403
9	Holding_Mobile_Phone	Teacher	385
10	Holding_Stick	Teacher	319
11	Reading_Book	Student	593
12	Sitting_on_Chair	Teacher	329
13	Sitting_on_Desk	Student	263
14	Slapping	Teacher	384
15	Sleeping	Student	219
16	Standing	Student	453
17	Talking	Student	294
18	Walking_in_Classroom	Teacher	403
19	Writing_on_Board	Teacher	419
20	Writing_on_Textbook	Student	548

We split the complete dataset into training, testing, and validation sets with proportions of 70%, 20%, and 10%, respectively. Figure 5 provides a glimpse of the action clips in the EduNet dataset.



Figure 5. A glimpse of EduNet dataset (Action categories from left to right: Arguing, Eating_in_Classroom, Holding_Book, Holding_Stick, Holding_Mobile_Phone, Explaining_the_Subject, Writing_on_Board, sleeping, Reading_Book).

4.2. Experimental Setup on the PARAM Shivay Supercomputer

The PARAM Shivay supercomputer is among the series of supercomputers designed and developed by the Centre for Development of Advanced Computing (C-DAC)

(www.cdac.in/, accessed on 2 November 2021), India, and the first supercomputer designed under the National Supercomputing Mission (NSM) (https://nsmindia.in/, accessed on 13 November 2021) and hosted at the Indian Institute of Technology (BHU), Varanasi, India. This supercomputer is based on Intel Xeon SKL G-6148, NVIDIA Tesla V100, with a total peak performance of 837 TFLOPS. The cluster consists of computing nodes connected with a Mellanox (ERD) InfiniBand interconnect network. The system uses the Lustre parallel file system. The GPU computing nodes are the nodes that have CPU cores along with accelerator cards. For some applications, GPUs attain a markedly high performance. To leverage this, we used CUDA libraries to map computations on the graphical processing units. Table 3 shows the configuration and Figure 6 illustrates the architecture diagram of the PARAM Shivay supercomputer.

Table 3. PARAM Shivay supercomputer configuration details.

Characteristics	Values
Peak Performance	837 Teraflops
Total number of nodes	233
Architecture	X86_64
Provisioning	xCAT 2.14.6
Cluster Manager	Openhpc (ohpc-xCAT 1.3.8)
Monitoring Tools	C-CHAKSHU, Nagios, Ganglia, XDMoD
Resource Manager	Slurm
I/O Services	Lustre Client
High-Speed Interconnects	Mellanox InfiniBand

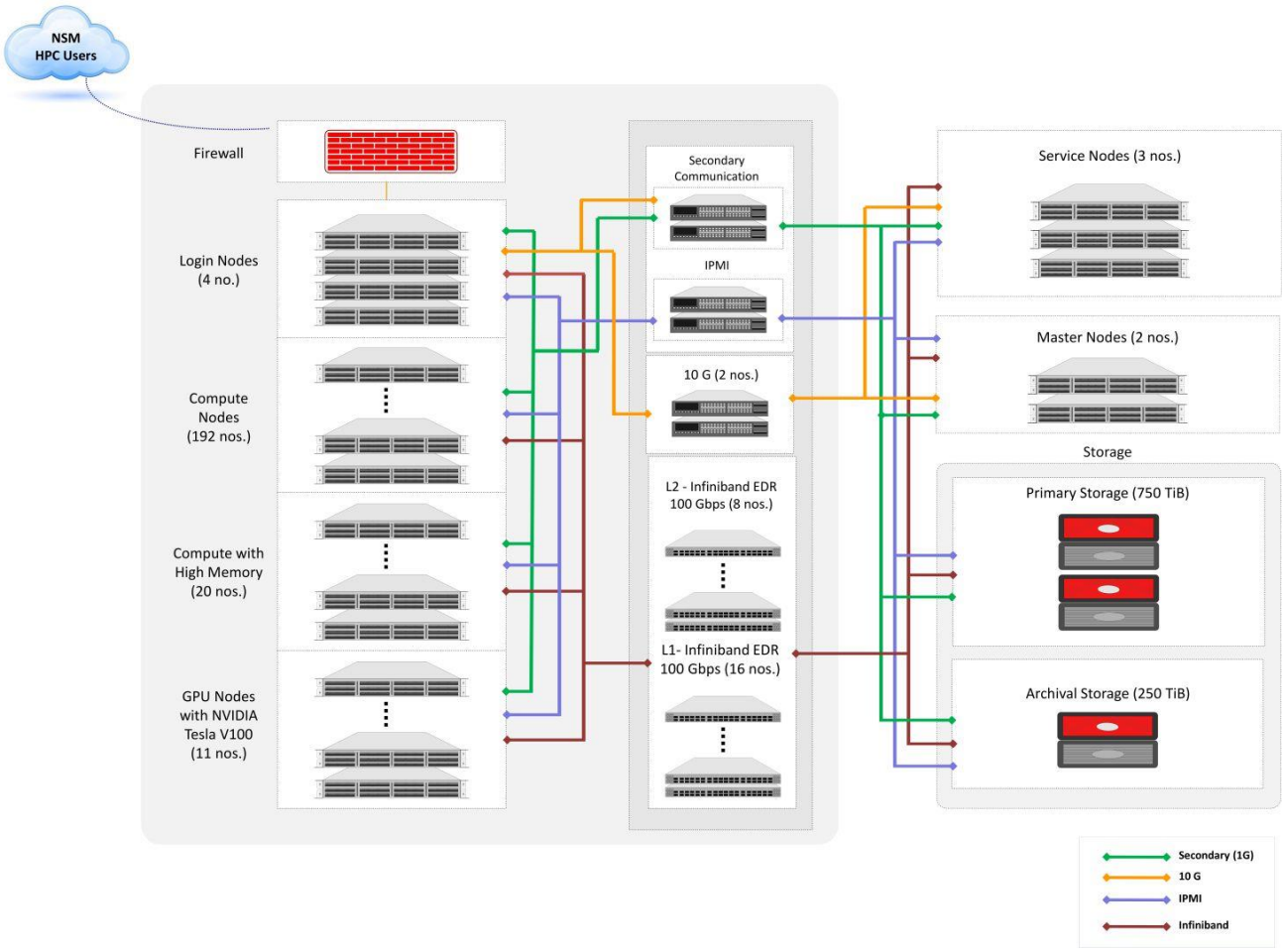


Figure 6. Architecture of the PARAM Shivay supercomputer [34].

In the proposed algorithms, the scenes of students and teachers are processed frame by frame in a parallel manner to classify them according to the model. At this stage, the model does not initially know which frame contains a student or a teacher. This information is determined by the single-shot detector, which recognises the scene based on the bounding box object. The dataset used in this method contains actions of both students and teachers. Therefore, parallel processing of both actions lead to action classification for either students or teachers.

4.3. Architecture of PARAM Shivay Supercomputer

Figure 6 shows the architecture diagram of the PARAM Shivay supercomputer.

All experiments in our proposed work are based on the CentOS 8.2 system, the CPU Intel Xeon E5-2620v4 (designed and developed by Intel, Santa Clara, CA, USA), and the V100 GPU. The development tool is PyCharm version 2.0, the programming language is Python 3, the image processing library is OpenCV4.1.1, and the deep-learning framework is TensorFlow 2.4 and Keras.

PARAM Shivay extensively uses modules, which serve to establish the production environment for a specific application independently of the application. These modules also determine which version of the application is accessible for a particular session. All applications and libraries are accessed through module files, and the user must load the appropriate module from those available modules. In a sample batch job submission script, we have listed the available versions of TensorFlow, after the TensorFlow GPU version with Python 3.6 is loaded in the working environment.

Peak Performance

The PARAM Shivay supercomputer consists of 2 master nodes, 4 login nodes, 4 service nodes, and 223 (CPU+GPU) nodes with a total peak computing capacity of 837 (CPU+GPU) TFLOPS performance.

Job Scheduler

In PARAM Shivay, the SLURM job scheduler is installed. SLURM (simple Linux utility for resource management) is a workload manager that provides a framework for job queues, allocation of computing nodes, and the start and execution of jobs.

Running deep-learning applications on PARAM Shivay

A sample batch job submission script

```
################################################

#!/bin/bash
#SBATCH -N 1
#SBATCH --ntasks-per-node=64
#SBATCH --gres=gpu:2
#SBATCH --error=job.%J.err
#SBATCH --output=job.%J.out
#SBATCH --time=01:00:00
#SBATCH --partition=gpu
// To load the module //
source conda/bin/activate
module load tensorflow 2.4
cd<Path of the executable>.
Python3 /home/myenv/run.py
################################################
```

4.4. Training and Validation on the EduNet Dataset

In our proposed work, the model utilised the mini-batch gradient descent (MBGD) method during training, where a mini-batch refers to a random selection of a training subset from the training dataset (assuming that a certain training set contains samples and the mini-batches). Running a mini-batch sample is usually considered a step towards the

completion of training. When all the running steps were completed, then the entire dataset was scrambled, and the above steps were repeated until the loss of the model converged or reached a satisfactory accuracy. If a training set contains n samples and the mini-batch size is b, then the entire dataset was divided into b×n mini-batches. The entire dataset was divided into sizes of 10.

The loss function for model training was categorical cross entropy (CE). The initial learning rate was 0.001, and the learning rate was adjusted by exponential decay, meaning that the decay factor was 0.1 and the decay step was set to 2000. The model's optimiser was "Adam" and the early stopping threshold was set to 10. The total training time was 12 h and 200 iterations. The loss curve achieved the value of 0.16, and an accuracy 96% for 80 epochs.

5. Result and Discussion

The proposed algorithm for recognising all 20 types of actions performed by students and teachers achieves an average recognition rate of 83.5% on the EduNet dataset and also outperforms some of the single-person actions such as Walking_in_Classroom and Standing.

Therefore, we conclude that the proposed algorithm of action recognition for students and teachers using the STAR-3D algorithm advances the recognition performance in an actual classroom environment. Figure 7 shows the validation accuracy of the STAR-3D algorithm on 20 action categories of the EduNet dataset, while Figure 8 shows the confusion matrix of the algorithm.

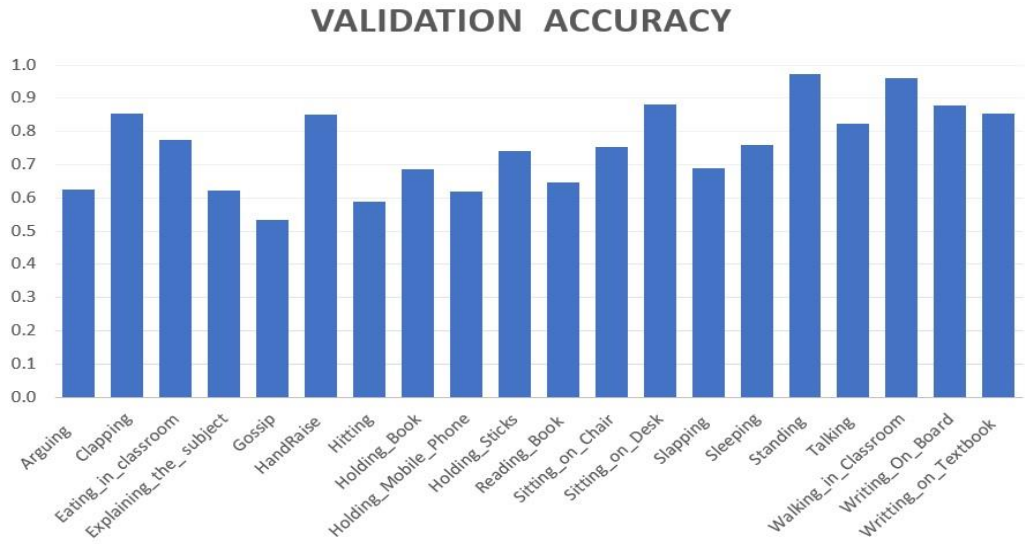


Figure 7. Validation accuracy of EduNet dataset action classes using STAR-3D.

STAR-3D demonstrates high accuracy in recognising single-person activities. Activities such as Standing, and Walking_in_Classroom, achieved the highest accuracy of 96.1% and 97.2% respectively, followed by Sitting_on_Desk (88.1%), Writing_on_Board (87.9%), Writing_on_Textbook (85.3%), Clapping (85.3%), and Hand_Raise (85.1%). The confusion matrix for validation of the STAR-3D algorithm on the EduNet dataset is shown in Figure 8. The student activity Gossip achieved the lowest recognition rate (53.4%), since multi-person student activity has a lot of distractions due to the combination of multiple activities. This is the limitation of STAR-3D, where single-person activity is recognised at a high accuracy rate, whereas multi-person activity recognition is at a low accuracy rate. Table 4 shows the comparative result of existing action recognition algorithms and STAR-3D.

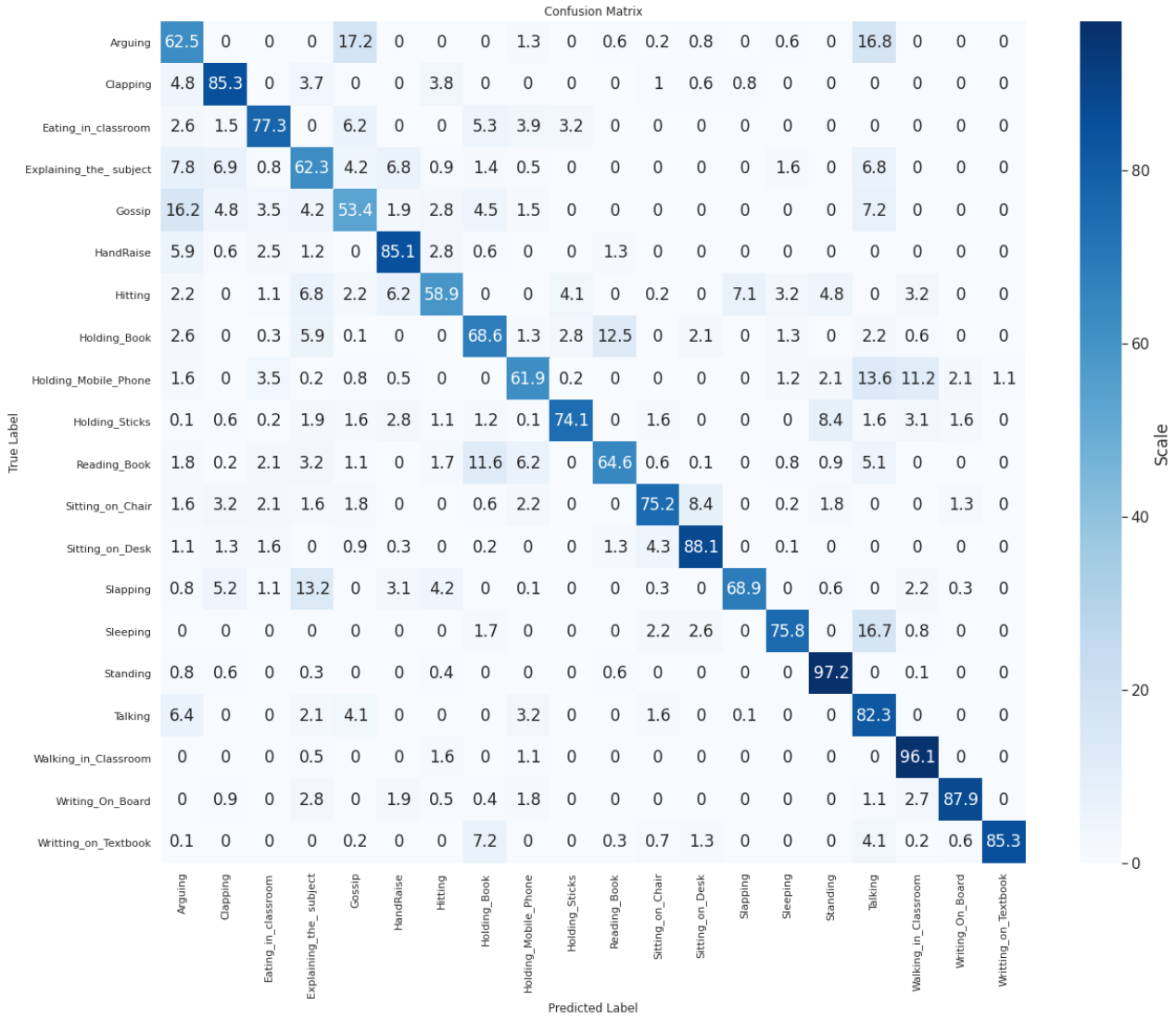


Figure 8. Confusion matrix.

Table 4. Comparison of proposed STAR-3D algorithm with existing algorithms.

Method	Purpose	Dataset	Pretrained On	Accuracy
Two-Stream I3D [12]	Human daily action recognition	UCF101, HMDB51	Kinetics	97.8%, 80.9%
3D BP-TBR [14]	Teacher behaviour recognition	TAD08 (Self-developed)	-	81.0%
		UCF101, HMDB51	Kinetics	97.11%, 81.0%
Feature Fusion Network [13]	Student behaviour recognition	UCF101, HMDB51	-	92.4%, 83.9%
STAR-3D	Action recognition of student and teacher	EduNet	-	83.5%

The 3D BP-TBR algorithm is pretrained on the Kinetics benchmark dataset and achieved a 97.11%, and 81.0% accuracy on the UCF101 and HMDB51 datasets, respectively, whereas it achieved 81.0% on the self-developed dataset. STAR-3D does not use any pre-trained data and shows an 82.5% accuracy on the EduNet dataset. We have also analysed that action recognition from the multi-person scene had some limitations; therefore in the activities such as Arguing, where 2–3 students are arguing with each other, our algorithm achieved a lower accuracy.

These methods for analysing student–teacher actions are applicable in the research areas focusing on evaluating interactive classroom by understanding the engagement

level of students and teachers. Human action recognition using deep-learning models is particularly beneficial for recognising actions in varying lighting conditions, background clutter, and different camera viewpoints. The proposed algorithm has the capacity to capture complex patterns and variations, making it suitable for the classroom and other learning environments.

From our experiments, several potential actions mapped into broad categories for the action recognition to measure the engagement level in the classroom environment:

Hand_Raise: Detection of students raising their hands to ask questions or to participate in discussions indicates their active participation.
Sleeping: Identifying students sleeping in the classroom indicates boredom and lack of participation.
Arguing: Recognising students arguing helps to understand their behaviour during interactions.
Eating_in_Classroom: Recognising students eating in the classroom indicates a lack of participation.
Explaining_the_Subject: Analysing how teachers respond to student questions or participation.
Walking_in_Classroom: Monitoring how teachers manage the classroom environment, including attention to students.
Holding_Mobile_Phone: Indicates that the teacher is distracted and not engaged with students during the class session.

Since HAR involves using AI algorithms to analyse and interpret human actions and movements, in the context of classroom engagement levels, it includes recognising students' and teachers' overall body language. These actions are indicative of the engagement level of students and teachers in a classroom environment. STAR-3D can be utilised in such scenarios to help educators to adopt a more interactive teaching approach, leading to increased engagement and motivation in the classroom, as well as improved learning outcomes.

Existing studies [18,20] reported student behaviour recognition in the classroom environment which can be applied to the understanding of student engagement level. Similarly, Gang et al. [14] explored teachers' behaviour recognition which helps towards the understanding of teachers' interaction. STAR-3D can recognise both the students' and teachers' actions, enabling a more comprehensive evaluation of the engagement level from both students' and teachers' perspectives.

6. Conclusions, Limitation and Future Research

We developed a student–teacher activity recognition model based on 3D CNN (STAR-3D) to identify human action in the classroom environment. This model is trained, tested, and validated on a self-built EduNet dataset which consists of 20 action categories of teachers and students. To handle the computational demands of video data, we trained the model on an HPC system with two GPUs, significantly speeding up the training time compared to a single GPU or desktop machine. In the scene-based recognition method, the model first identifies the "student scene" and "teacher scene" using the SSD sampling method enhanced by GAN and validates them with an action dictionary. Thereafter, the input video and optical flow are applied to a 3D CNN (with the base model ResNet50) in a spatiotemporal manner with an SVM classifier in the last layer. To enhance performance, we used "L1" regularisation and the Kalman filter. Our model efficiently identifies the action with the best performance of 83.5%.

While our research achieved its objectives, there are limitations to consider for future work. To begin with, the proposed algorithm has only been tested and validated on the EduNet dataset and not on other benchmark datasets. Additionally, actions are currently limited to 20 categories.

In the future work, we will focus on more accurate and elaborate identification of student–teacher activities with a greater number of categories. The future scope will have the potential for recognising a wider range of human actions in the classroom environments such as exam monitoring, student presentation, teachers' feedback, behaviour analysis,

inclusive education, students with special needs, etc. A different model such as I3D can also be applied in an experimental way to test the improvement in model accuracy, pre-trained with benchmark datasets such as UCF101, HMDB51, and Kinetics. Furthermore, the STAR-3D model can also be applied and compared with the UCF101, HMDB51, and Kinetics datasets apart from the EduNet dataset. Transformer-based action recognition, which is gaining popularity, also offers opportunities for improving the model accuracy. Integrating the model with humanoid robots for classroom monitoring and student engagement is another area to be explored. This would enable a comparative study of action recognition with and without robot assistance.

Author Contributions: Conceptualisation, V.S., M.G. and A.K.; methodology, V.S. and M.G.; software, V.S.; validation, M.G., A.K. and D.M.; formal analysis, M.G. and D.M.; investigation, M.G.; resources, V.S.; data curation, V.S.; writing—original draft preparation, V.S.; writing—review and editing, V.S. and M.G.; visualisation, V.S. and M.G.; supervision, M.G. and A.K.; project administration, A.K.; funding acquisition, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The EduNet dataset used for this study can be found on this website https://edunet-drsta.com/, accessed on 22 October 2021 by filling the access form and agreeing upon the terms and conditions. This dataset will be provided only for research purposes.

Acknowledgments: We acknowledge the National Supercomputing Mission (NSM) for providing computing resources of 'PARAM Shivay' at the Indian Institute of Technology (BHU), Varanasi, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Brandisauskiene, A.; Cesnaviciene, J.; Bruzgeleviciene, R.; Nedzinskaite-Maciuniene, R. Connections between teachers' motivational behaviour and school student engagement. *Electron. J. Res. Educ. Psychol.* **2021**, *19*, 165–184. [CrossRef]
2. Obenza-Tanudtanud, D.M.; Obenza, B. Evaluating Teacher-Student Interaction and Student Learning Engagement in the New Normal: A Convergent-Parallel Design. *Psychol. Educ. A Multidiscip. J.* **2023**, *15*, 1–13.
3. Kundu, A.; Bej, T.; Dey, K.N. Time to grow efficacious: Effect of teacher efficacy on students' classroom engagement. *SN Soc. Sci.* **2021**, *1*, 266. [CrossRef]
4. Pabba, C.; Kumar, P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Syst.* **2022**, *39*, e12839. [CrossRef]
5. Fannakhosrow, M.; Nourabadi, S.; Ngoc Huy, D.T.; Dinh Trung, N.; Tashtoush, M.A. A Comparative Study of Information and Communication Technology (ICT)-Based and Conventional Methods of Instruction on Learners' Academic Enthusiasm for L2 Learning. *Educ. Res. Int.* **2022**, *2022*, 5478088. [CrossRef]
6. Zhai, X.; Chu, X.; Chai, C.S.; Jong, M.S.Y.; Istenic, A.; Spector, M.; Liu, J.-B.; Yuan, J.; Li, Y. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity* **2021**, *2021*, 8812542. [CrossRef]
7. Miao, F.; Holmes, W.; Huang, R.; Zhang, H. *AI and Education: Guidance for Policy-Makers*; United Nations Educational, Scientific and Cultural Organization: Paris, France, 2021.
8. Whitehill, J.; Serpell, Z.; Lin, Y.-C.; Foster, A.; Movellan, J.R. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Trans. Affect. Comput.* **2014**, *5*, 86–98. [CrossRef]
9. Vanneste, P.; Oramas, J.; Verelst, T.; Tuytelaars, T.; Raes, A.; Depaepe, F.; Van den Noortgate, W. Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement. *Mathematics* **2021**, *9*, 287. [CrossRef]
10. Dimitriadou, E.; Lanitis, A. Student Action Recognition for Improving Teacher Feedback during Tele-Education. *IEEE Trans. Learn. Technol.* **2023**, *17*, 569–584. [CrossRef]
11. Bourguet, M.-L.; Jin, Y.; Shi, Y.; Chen, Y.; Rincon-Ardila, L.; Venture, G. Social robots that can sense and improve student engagement. In Proceedings of the 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Takamatsu, Japan, 8–11 December 2020; pp. 127–134.
12. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
13. Jisi, A.; Yin, S. A new feature fusion network for student behavior recognition in education. *J. Appl. Sci. Eng.* **2021**, *24*, 133–140.

14. Gang, Z.; Wenjuan, Z.; Biling, H.; Jie, C.; Hui, H.; Qing, X. A simple teacher behavior recognition method for massive teaching videos based on teacher set. *Appl. Intell.* **2021**, *51*, 8828–8849. [CrossRef]
15. Sharma, V. Deep Learning for Human Action Recognition in the Classroom Environment. Ph.D. Thesis, Banaras Hindu University, Varanasi, India, 2021.
16. Chang, M.-J.; Hsieh, J.-T.; Fang, C.-Y.; Chen, S.-W. A vision-based human action recognition system for moving cameras through deep learning. In Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning, Hangzhou, China, 27–29 November 2019; pp. 85–91.
17. Nida, N.; Yousaf, M.H.; Irtaza, A.; Velastin, S.A. Instructor activity recognition through deep spatiotemporal features and feedforward extreme learning machines. *Math. Probl. Eng.* **2019**, *2019*, 2474865. [CrossRef]
18. Zhang, R.; Ni, B. Learning behavior recognition and analysis by using 3D convolutional neural networks. In Proceedings of the 2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST), Luang Prabang, Laos, 2–5 July 2019; pp. 1–4.
19. Li, X.; Wang, M.; Zeng, W.; Lu, W. A students' action recognition database in smart classroom. In Proceedings of the 2019 14th International Conference on Computer Science & Education (ICCSE), Toronto, ON, Canada, 19–21 August 2019; pp. 523–527.
20. Cheng, Y.; Dai, Z.; Ji, Y.; Li, S.; Jia, Z.; Hirota, K.; Dai, Y. Student action recognition based on deep convolutional generative adversarial network. In Proceedings of the 2020 Chinese Control And Decision Conference (CCDC), Hefei, China, 22–24 August 2020; pp. 128–133.
21. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. Available online: https://www.crcv.ucf.edu/data/UCF101.php (accessed on 12 November 2021).
22. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
23. Zuo, K.; Su, X. Three-Dimensional Action Recognition for Basketball Teaching Coupled with Deep Neural Network. *Electronics* **2022**, *11*, 3797. [CrossRef]
24. Qiu, Q.; Wang, T.; Chen, F.; Wang, C. LD-Recognition: Classroom Action Recognition Based on Passive RFID. *IEEE Trans. Comput. Soc. Syst.* **2023**, *11*, 1182–1191. [CrossRef]
25. Ren, H.; Xu, G. Human action recognition in smart classroom. In Proceedings of the Fifth IEEE International Conference on Automatic Face Gesture Recognition, Washington, DC, USA, 21 May 2002; pp. 417–422.
26. Raza, A.; Yousaf, M.H.; Sial, H.A.; Raja, G. HMM-based scheme for smart instructor activity recognition in a lecture room environment. *SmartCR* **2015**, *5*, 578–590. [CrossRef]
27. Sharma, V.; Gupta, M.; Kumar, A.; Mishra, D. EduNet: A new video dataset for understanding human activity in the classroom environment. *Sensors* **2021**, *21*, 5699. [CrossRef] [PubMed]
28. Li, X.; Hua, Z.; Li, J. Attention-based adaptive feature selection for multi-stage image dehazing. *Vis. Comput.* **2023**, *39*, 663–678. [CrossRef]
29. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14, 2016. pp. 21–37.
30. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 139–144.
31. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
32. Pei, Y.; Biswas, S.; Fussell, D.S.; Pingali, K. An elementary introduction to Kalman filtering. *Commun. ACM* **2019**, *62*, 122–133. [CrossRef]
33. Ajagekar, A. Adam. Available online: https://optimization.cbe.cornell.edu/index.php?title=Adam (accessed on 15 November 2021).
34. India, C.-D. PARAM SHIVAY Architecture Diagram. 2019. Available online: https://www.iitbhu.ac.in/cf/scc/param_shivay/architecture (accessed on 13 November 2021).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.