Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 1, April 2026, pp. 149∼163
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i1.pp149-163

Student activity recognition from classroom video: a survey

Phuong-Dung Nguyen[1,2,3], Khanh-Huyen Bui[1,2], Thi-Lan Le[1,2]
[1]School of Electrical and Electronic Engineering (SEEE), Hanoi University of Science and Technology (HUST), Hanoi, Vietnam
[2]SigM Laboratory, SEEE, Hanoi University of Science and Technology (HUST), Hanoi, Vietnam
[3]Thuyloi University, Hanoi, Vietnam

Article Info

Article history:

Received Apr 3, 2025
Revised Feb 20, 2026
Accepted Mar 4, 2026

Keywords:

Activity detection
Classroom video
Student activity recognition

ABSTRACT

Student behavior and activity play a crucial role in shaping the classroom atmosphere and influencing the quality of a learning session. Recently, vision-based student activity recognition has gained significant attention. However, recognizing student activities from classroom videos presents unique challenges due to the nature of the classroom environment, such as the presence of multiple students and severe occlusions. As a result, research in this area has often overlooked these challenges. This study provides a detailed and comprehensive review of student activity recognition from classroom videos. First, we formalize the problem of student activity recognition from videos and categorize existing methods into three distinct approaches: frame-level, clip-level, and continuous recognition. We then provide a detailed analysis of representative methods for each approach. In addition, we present a comprehensive overview of publicly available datasets for student activity recognition and discuss key open challenges, together with potential future research directions. Our analysis reveals that: (1) Most existing studies focus on frame-level recognition, while clip-based and continuous activity recognition remain relatively underexplored; (2) there is still a lack of large-scale, standardized benchmark datasets for vision-based student activity recognition; and (3) existing research primarily emphasizes recognition accuracy, whereas real-time performance and computational efficiency are rarely addressed.

Corresponding Author:

Thi-Lan Le
School of Electrical and Electronic Engineering (SEEE)
Hanoi University of Science and Technology (HUST)
Dai Co Viet, Hanoi, Vietnam
Email: lan.lethi1@hust.edu.vn

1. INTRODUCTION

Activity recognition involves monitoring and analyzing human behavior and the surrounding environment to identify or infer ongoing activities [1]. The main objective of activity recognition is to provide insights into users' activities, states, and behaviors, enabling proactive computational systems to offer personalized assistance and support. This process often relies on data collected from sensors such as cameras, wearable devices, and other information sources. In recent years, activity recognition has gained significant attention from the research community thanks to its broad applications in human-computer interaction, abnormal activity detection for smart cities, and gesture assessment in physical therapy for patients, etc.

In the educational domain, student behavior and activity play a crucial role in shaping the classroom atmosphere and determining the quality of a learning session. Conventional methodology based on teacher observations, though commonly practiced, is often subjective and faces challenges in monitoring all students,

particularly in large classrooms or when visual obstructions occur. To address these challenges, analyzing visual [2], auditory [3], or physical signals [4] presents promising solutions for automatic classroom assessment. Among these modalities, the visual modality is the most widely used, as cameras are often readily available in classrooms and can provide rich contextual information for more accurate student activity recognition.

Student activity recognition from classroom videos an be considered as a sub topic of human activity recognition (HAR) howwever it presents specific challenges. The first challenge is severe occlusion. Students may be obscured by other students or by classroom furniture such as chairs and desks, complicating accurate detection and tracking of individuals. The second challenge is the high number of subjects (i.e., students) in the scene. Most activity recognition methods assume the presence of only one or two subjects. However, in a classroom context, the number of students is usually high, with different students potentially engaging in various activities at different times.

While numerous surveys have been conducted on HAR in general [5]-[8], and within specific domains such as sports [9], [10], there is a lack of surveys focusing specifically on student activity recognition [11]. This study focuses on student activity recognition from classroom videos. Specifically, we present a detailed and comprehensive review of existing methods, including those that rely solely on spatial information and those incorporating both spatial and temporal features. Additionally, we provide a detailed overview of datasets collected for student activity recognition. Furthermore, we discuss various open challenges and propose potential future research directions. The key contributions of this paper are as follows: (1) We define the problem of student activity recognition and categorize it into three distinct approaches; (2) We present a state-of-the-art review of methods and datasets used for student activity recognition in classroom settings; (3) We identify open challenges and suggest possible directions for future research.

2. STUDENT ACTIVITY RECOGNITION FORMULATION

Student activity recognition can be formally defined as follows: Given an untrimmed video with T consecutive frames, student activity recognition aims to detect a set of $\mathbf{P}$ of M tubelets P_i, each corresponding to an instance of an activity performed by a student. The set of tubelets is represented as: $\mathbf{P} = \{P_1, P_2, P_i, \ldots, P_M\}$, where each tubelet P_i is defined as,

$$P_i = \left(\left\{P_i^t \mid t = x, x+1, \ldots, x+N_{P_i}\right\}, c_i\right) \tag{1}$$

with: P_i^t: is the bounding box at frame t of tubelet i^{th} that represents the spatial localization of the activity instance at frame t; x: is the starting frame where the activity instance occurs; N_{P_i} is the duration of the activity (number of frames in the tubelet); c_i: is the activity class label associated with the tubelet.

Figure 1 illustrates a recognized tubelet P_i of class c_i is raising hand. The starting frame of the tubelet is 0672 $x = \#0672$ whereas the length of tubelet is 133 ($N_{P_i} = 133$).

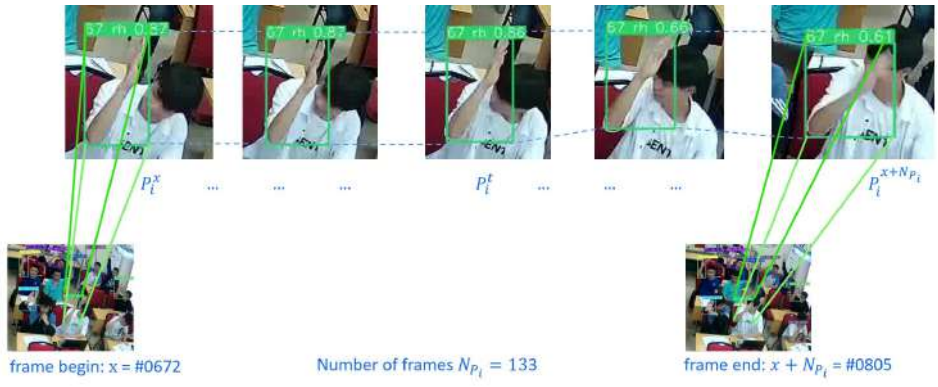

Figure 1. An example of a tubelet output for student activity recognition

Student activity recognition can be formally defined as a mapping from video frames to a set of tubelets. Given an untrimmed video with T consecutive frames, we define a function,

$$f : \{I_1, I_2, \ldots, I_T\} \rightarrow \mathbf{P} \tag{2}$$

Where: $\{I_t\}_{t=1}^{T}$ represents the sequence of video frames; f is a function that extracts and recognizes student activities; $\mathbf{P} = \{P_1, P_2, \ldots, P_M\}$ is the set of tubelets corresponding to detected student activities defined by (1). Thus, the function f maps a sequence of frames to a structured set of tubelets, capturing student activities in the classroom.

In recent years, several works have been dedicated to classroom activity recognition and encouraging results have been achieved. Some approaches rely solely on spatial information, while others integrate both spatial and temporal data. Based on the type of information used, existing methods for student activity recognition can be categorized into three approaches: Frame-level, clip-level, and continuous activity recognition. Frame-level methods aim to detect instances of activities of interest within individual image frames. The majority of methods for student activity recognition fall into the frame-level category thanks to the emergence and advancements in object detection techniques [12]-[21]. Methods of the clip-level approach classify pre-segmented video clips into specific classes. It is worth noting that the term *clips* in student activity recognition refers to a sequence of regions (i.e., bounding boxes) of the activity in the original frames. In the activity recognition field, a clip may refer to a sequence of entire frames, assuming that each frame contains only one person. However, this is not the case for student activity recognition. Although some methods belonging the clip-level approach have been proposed [22], [23], these methods have drawback because in a classroom setting, where many students may be performing different activities at different times, determining the appropriate clips is not a straightforward task. Continuous student activity recognition, which aims to determine the location of the activity instance and track its changes over time, is the most suitable approach. However, due to the challenges of student activity recognition, very few works have been successfully developed for this approach [24].

It is worth noting that in the (1), the frame-level methods can only determine the individual bounding boxes P_i^t for each type of activity. However, they do not provide information on whether the detected bounding boxes across frames belong to the same activity instance. The clip-level approach can determine only the class c_i for a given tubelet P_i, which is predefined. In contrast, continuous activity recognition methods enable the determination of both the tubelet P_i and its corresponding class c_i. In the following sections, the methods for student activity recognition in each category will be analyzed.

3. METHODS

This section outlines the procedure used to collect, screen, and categorize the article for a comprehensive survey of student activity recognition methods in classroom videos. The overall methodology is adapted from the systematic review framework presented in the study [25], and refined to align with the three methodological categories of this study: frame-level, clip-level, and continuous recognition. The workflow consists of four major stages-Identification, Screening, Eligibility, and Included-corresponding to the process illustrated in Figure 2. The entire workflow is designed to accurately reflect the progression of research in the period 2013–2025 and to ensure consistency between the collected sources and the analytical structure presented in the subsequent sections.

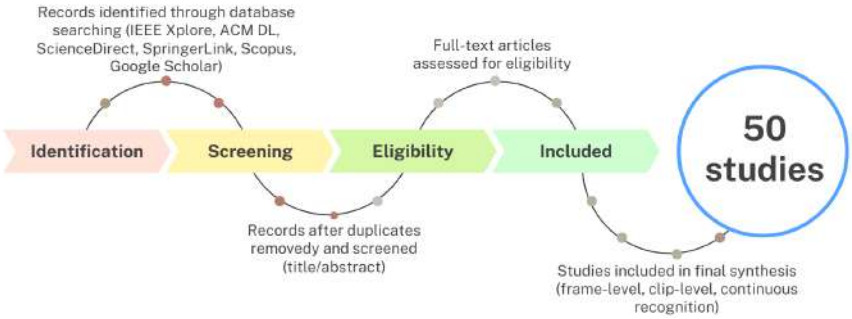

Figure 2. Four-stage article selection workflow: Identification, screening, eligibility, and included

3.1. Step 1: Identification

Based on the scope of this survey, we conducted a systematic search across several reputable academic databases, including IEEE Xplore, ACM Digital Library, Elsevier ScienceDirect, SpringerLink, Scopus, and

Google Scholar. The search targeted studies directly related to student activity recognition from classroom videos, using two main groups of keywords,

- Keywords related to the target domain "student activity recognition", "classroom activity recognition", "student behavior analysis", "classroom video", "hand-raising detection".
- Keywords related to technical approaches "object detection", "pose estimation", "spatio-temporal action detection", "skeleton-based action recognition", "clip-level activity recognition", "tracking", "Simple online and real-time tracking (SORT)", "tubelet detection".

3.2. Step 2: Screening

The screening stage was carried out in two sequential filtering steps to remove studies that were clearly irrelevant before conducting full-text assessment.

- Title Screening: The titles of all retrieved publications were examined to quickly eliminate studies that did not align with the research scope. Titles indicating a focus on generic HAR, human motion analysis without educational relevance, or applications unrelated to classroom environments were excluded.
- Abstract Review: The goal was to determine whether each study explicitly addressed student-centered behaviors observable in classroom video recordings. Studies were excluded if the abstract did not clearly identify student activities as the primary target, if the behavioral categories were ambiguous, or if the methodological focus did not involve recognition or analysis of student actions.

Only studies that satisfied both criteria-relevance in title and clarity in abstract-were retained for the eligibility stage.

3.3. Step 3: Eligibility

The eligibility criteria were established to ensure that only studies with direct and meaningful contributions to the problem of student activity recognition in classroom settings were included in this survey.

- Inclusion criteria:
  - Published within the period 2013–2025.
  - Appeared in peer-reviewed journals, international conferences, or book chapters.
  - Focused on student activity recognition in classroom environments, at one of the three levels: frame-level, clip-level, or continuous-level.
  - Provided a complete description of the methodology, model architecture, processing pipeline, and experimental protocol.
  - Reported quantitative evaluation metrics enabling performance comparison (e.g., accuracy, mAP, F1-score).
- Exclusion criteria:
  - Studies addressing general HAR without a classroom context.
  - System-description papers that lack model evaluation.
  - Works for which the full text is not available.
  - Non-academic materials such as editorials, keynote talks, opinion pieces, or slide presentations.
  - Studies that do not report or clearly define bounding boxes, tubelets, clips, or action labels, making analysis and comparison infeasible.

3.4. Step 4: Included (Final set of studies)

In the initial search phase, a total of 80 publications were retrieved from the selected databases. Through a multi-stage screening process, studies that did not fall within the scope of student activity recognition, did not clearly describe classroom behaviors, or did not address activities such as hand-raising, sleeping, standing, or phone usage were excluded. Additional studies were removed due to insufficient methodological detail, lack of quantitative results, or misalignment with the three target methodological categories (frame-level, clip-level, and continuous recognition). Ultimately, 50 studies met all eligibility criteria and were retained for in-depth analysis. These studies were then categorized into the three main methodological groups defined in this survey: frame-level, clip-level, and continuous recognition.

4. RESULTS AND DISCUSSIONS

4.1. Frame-level approach for student activity recognition

Frame-based methods focus on recognizing target activities within single image frames. This approach is closely related to object detection in still images, a research area that has attracted significant attention in computer vision and machine learning. As shown in Figure 3, these activity recognition methods take individual frames as input and generate corresponding outputs. Specifically, for each frame, the models produce bounding boxes that locate the relevant activities. While some approaches are designed to identify only one type of activity, others are capable of detecting multiple activities at the same time. Owing to the rapid progress in object detection techniques, a wide range of frame-based methods for student activity recognition has been developed.

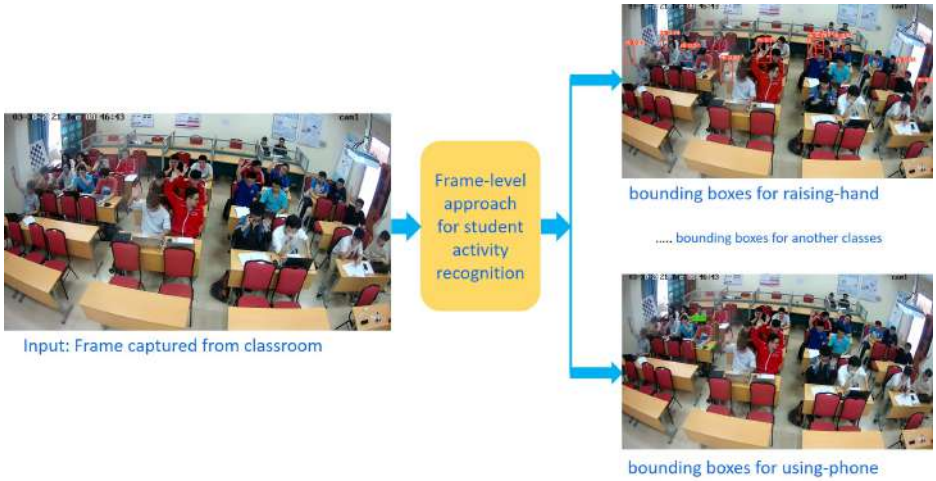

Figure 3. Frame-level approach for student activity recognition

Among student activities, hand-raising has attracted the most attention from the research community as the number of hand-raising can be used to infer the quality of teaching session. However, detecting hand-raising from images is challenging due to the occluded body parts and low-resolution hands areas. In [12], the authors proposed to apply Lienhart-Maydt method for hand-raising detection. Then, the authors proposed to combine the face detection results with those of hand detection within the same frame to enhance the detection accuracy. When both face and hand are detected, the relative positions between them are calculated to determine if the person is actually raising their hand. The work in [13] proposed a model to detect the gesture of raising a hand using static images as input. Initially, the model selects regions of interest likely to contain the gesture of raising a hand, then uses HSV and YCbCr color space models to detect skin color in the images. The edges are detected using a stable Gaussian filter, and finally, the activity is classified as raising a hand or not.

With the arrival of deep-learning based object models, recent works mainly rely on these detection models to detect hand raising from classroom images. Deep learning models for object detection are generally divided into two main types: two-stage object detectors (e.g. Region-based fully convolutional network (R-FCN) [26], Faster region-based convolutional neural network (RCNN) [27]) and one-stage object detectors (e.g. You only look once (YOLO) and its variants [28]).

The authors in [29] proposed an R-FCN-based method for detecting hand-raising gestures in images captured within a real-world classroom setting. To enhance the detection of hand-raising gestures, even in low-resolution images, they integrated pyramid features into the original R-FCN architecture. Specifically, the R-FCN network was improved by incorporating pyramid features into its design.

Relying solely on object detection for hand-raising recognition may lead to false alarms and missed detections. To address this issue, additional cues such as hand pose information have been incorporated. In [15], the authors proposed a three-step framework consisting of hand-raising detection, pose estimation, and matching steps. In the first step, an improved R-FCN algorithm was applied for hand detection. In the second step, a novel part affinity fields-based (PAF) pose estimation method was introduced to detect human body

keypoints. Finally, a heuristic matching strategy based on the spatial relationship between detected hands and body keypoints was employed to identify students who were raising their hands. Similarly, [18] presented an approach for recognizing hand-raising gestures consisting of two stages. The first stage uses multi-stage pose estimation to determine candidate hand regions for each student. Then, the second stage applies a binary classification network to determine the specific gesture type. Both studies relied on the pose estimation technique. However, this technique only worked well in simple scenes without much occlusion or clutter.

An extended non-local module combined with the Libra-RCNN detection model was proposed in study [30], enhancing image information extraction by incorporating local context and spatial correlation between pixels. The model addressed two main limitations: the lack of locality-awareness and the disregard for spatial relationships between pixels.

Recently, [31] proposed a novel machine learning approach that is robust to viewpoint variations and occlusions. The method leverages long short-term memory (LSTM) networks to detect hand-raising activity from pose estimation's information. Similarly, the authors in [32] proposed a morphology-based analysis method. The proposed method in this work utilized YOLOx for object detection and HrNet for skeleton estimation. Students' skeleton key point data is converted into several one-dimensional time series, allowing for a detailed analysis of hand-raising behaviors. These models help in accurately capturing and analyzing hand-raising actions.

Following a similar approach to improving R-FCN, the study [17] detected sleep persons in classrooms by leveraging pyramid feature representation. They employed a modified R-FCN model integrated a feature pyramid and deformable convolution to solve the challenges in sleep gesture detection, such as occlusion and diversity of gestures. By integrating both together, the system linked feature layers at various scales. This made the proposed method more effective in detecting small sleep gestures. Additionally, the study proposed using deformable convolution networks combined with local multiscale testing, enabling the system to learn the specific characteristics of small-size sleep gestures, such as various postures and changes in the bounding box sizes.

Some works tried to recognize more activities. In [33], the authors proposed a method based on CNNs combined with transfer learning. Pre-trained architectures, such as VGG16 and VGG19, were utilized to extract deep features from images. In [19], the authors developed an intelligent system capable of automatically recognizing student behaviors in recorded classrooms including raising hands, standing up, and dozing off. These behaviors are challenging to detect due to scale variations, low resolution, and imbalanced sample distributions. They enhanced Faster R-CNN by introducing a new scale-aware detection head to handle scale variations, a feature fusion strategy for detecting low-resolution behaviors with minimal extra computation, and Online hard example mining (OHEM) to mitigate severe class imbalances. Additionally, they proposed a technique that combines feature vectors from different layers to create the feature map for detection, improving the accuracy of detecting small objects, such as sleeping and hand-raising gestures.

The study [20] introduced a new detection model called GestureDet. This model allows for the detection of typical student gestures, including raising hands, standing up to speak, and sleeping. The authors enhanced the MobileNetV2 object detection model by integrating spatial attention, channel attention, and batch attention mechanisms to learn features more robustly from data. Additionally, GestureDet's lightweight nature allows to deploy it on embedded devices such as the NVidia Jetson TX2.

The study [34] focuses on the task of recognizing seven common types of student activities in the classroom by leveraging the power of pre-trained CNN models. The authors apply transfer learning techniques to fine-tune well-known network architectures such as VGG-16, ResNet-50, Inception V3, and Xception on a self-constructed classroom dataset consisting of more than 4,000 images.

The study on ET-YOLOv5s [35] addressed the challenges posed by low-resolution classroom environments and small-scale objects by incorporating an enhanced super-resolution generative adversarial network (ESRGAN). The ESRGAN module was employed to enhance image details prior to inputting the images into the detection network.

In 2024, Jia and He [36] proposed the SBD model, a hybrid framework that integrates YOLOv5 with the coordinate attention (CA) mechanism to generate high-quality feature maps, replacing the traditional VGG-19 backbone in the OpenPose framework. The proposed model not only performs object detection but also facilitates detailed analysis of human keypoints, thereby improving behavior recognition performance under conditions of partial student occlusion.

The authors in [21] developed a student behavior recognition system based on skeleton pose estimation

and person detection to assess students' learning attitude. Four main student activities are considered including looking, asking, boring and bowing. Based on the analysis results, this system can evaluate student attitudes towards the lesson, providing useful information for teachers and education managers. In [2], the authors evaluated the performance of three object detection models including Faster R-CNN, YOLOv5, and detection transformer (DETR) for student activity recognition on the StudentAct dataset.

Table 1 summarizes the methods that follow the frame-level approach. Although these methods have achieved promising results, they often suffer from missed detections and false alarms due to their inability to capture the temporal aspect of activities.

Table 1. Summary of frame-level methods for student activity recognition

Method	Description	Activities	Result	Level
Nazaré and Ponti (2013) [12]	Identify hand-raising posture by analyzing the positional relationship between the face and hands	Raising hand	Achieved over 60% accuracy on most of the their test data	Frame
Jesna *et al.* (2016) [13]	Use HSV and YCbCr color space for skin color detection and Gaussian filters for edge detection	Raising hand	Achieved 91% accuracy on their hand dataset	Frame
Zhou *et al.* (2018) [15]	Use Pose estimation to identify key body points and combine positional information	Raising hand	Achieved 83% accuracy on their dataset (30 schools)	Frame
Si *et al.* (2019) [29]	Enhance R-FCN by integrating the Feature Pyramid to detect hand-raising gestures	Raising hand	Achieved 90% mAP on self-constructed dataset	Frame
Liao *et al.* (2019) [18]	Utilize multi-stage pose estimation to identify the hand-based region	Raising hand	Achieved accuracy of 94.76%	Frame
Buhler *et al.* (2023) [31]	Use LSTM to detect hand-raising actions from pose estimation results	Raising hand	F1-score of 76%	Frame
Chen *et al.* (2024) [32]	Use YOLOx and HrNet for skeleton estimation, transformed into time series	Raising hand	Analysis of activity's speed and amplitude	Frame
Li *et al.* (2019) [17]	Modified R-FCN with feature pyramid and deformable convolution	Sleeping	Achieved 0.74 AP@0.5	Frame
Hoang *et al.* (2019) [33]	CNNs (VGG16, VGG19) combined with transfer learning	8 activities (writing, reading, etc)	VGG19 achieved 80.8% accuracy	Frame
Zheng *et al.* (2020b) [19]	Faster R-CNN with Scale-aware Detection Head and OHEM	Hand-raising, standing, sleeping	Achieved 57.6% mAP	Frame
Zheng *et al.* (2020a) [20]	GestureDet: Improved MobileNetV2 with spatial and channel attention	Hand-raising, standing, sleeping	Achieved 74.5% mAP	Frame
Lin *et al.* (2021) [21]	Skeleton pose estimation and human detection	Asking, boring, bowing, looking classes	Precision 89%, Recall 91%	Frame
Nguyen *et al.* (2022) [2]	Evaluate Faster R-CNN, YOLOv5 and DETR	Standing, sitting, phone, sleeping, hand	YOLOv5 achieved best mAP (up to 94.3%)	Frame
Lina *et al.* (2022) [35]	Integrate ESRGAN to restore image details	11 activities (bowing, drinking, etc)	Achieved a mAP of 96.8%	Frame
Deshpande and Deshpand (2023) [34]	Transfer learning with VGG-16, ResNet-50, Xception	7 activities (discussion, writing, etc)	Xception achieved 92% accuracy	Frame
Jia and He (2024) [36]	YOLOv5 with Coordinate Attention replacing VGG-19 in OpenPose	Raising hand, standing, writing, etc	Obtained an mAP of 82.1%	Frame

Among the studied activities, hand-raising has received the most attention from the research community. However, its detection performance varies significantly across different methods, ranging from 39.4% AP in [2] to 94.76% accuracy in [18]. A direct comparison of these methods is infeasible, as they are evaluated on private datasets collected by the respective authors. Additionally, other important activities, such as using a phone, have not been fully explored. Future research should focus on improving detection performance and extending recognition to a broader range of activities.

Besides using appearance features, some methods have attempted to integrate human skeleton information to reduce false alarms. However, previous studies were conducted in classroom environments with

minimal occlusion and clutter, where human joints could be accurately estimated. In real classroom conditions, joint estimation performance may be less reliable, posing additional challenges for activity recognition.

Finally, from a practical point of view in classroom management, it is crucial to recognize complete activity instances from beginning to the end rather than analyzing individual frames as recognizing full instances enables appropriate interventions. For example, detecting complete instances of negative activities, such as using a phone or sleeping, allows measurement of their duration and frequency. This information can help inform decisions to adjust classroom content to better engage the students.

4.2. Clip-level approach for student activity recognition

This approach focuses on classifying a predefined tubelet (i.e., clip) into an action class as illustrated in Figure 4. Once the tubelet is defined, the problem becomes similar to action recognition and classification. Therefore, methods developed for action recognition can be applied to clip-level student activity recognition [8], [37].

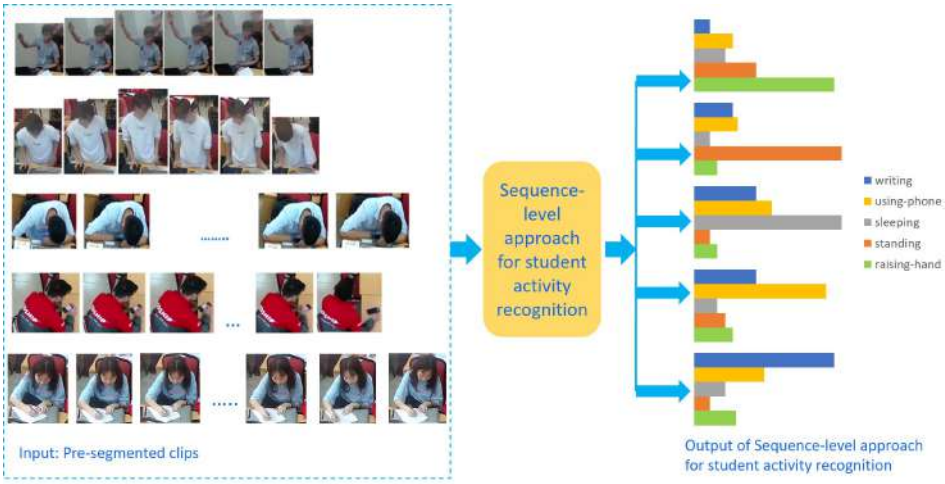

Figure 4. Clip-level activity recognition approaches

In [38], the authors proposed to employ a combination of traditional features, including local log-euclidean multivariate gaussian (L2EMG) and scale-invariant feature transform (SIFT) for student activity recognition. These features were used to capture shape and texture variations of students in video clips for the classification of five basic actions, namely raising hands, standing, sitting, writing, and reading.

In the study [39], the authors fine-tuned the two-stream I3D-ResNet-50 model [40] - an advanced 3D video classification architecture that uses a 3D convolution network to directly learn and extract spatiotemporal information from video data to evaluate the performance of the EduNet dataset, which consists of action data from classroom environments.

Based on the joint estimation algorithm, study [23] proposed a model for recognizing student activities. The model in the study combines the Alphapose joint estimator and the DD-Net activity recognition network, creating a compact deep learning model that still delivers high recognition accuracy. First, Alphapose extracts joint coordinates from classroom videos, dividing them into sequences from the start frame to the end frame of each activity for each student, and labels them accordingly. The input to DD-Net is the sequences of joint coordinates of a person performing a single activity.

Another clip-level study introduced in [22], based on the SlowFast network, incorporates the multi-scale spatial-temporal attention (MSTA) module into the Slow path, which includes extracting multi-scale spatial features, channel attention, and temporal attention to effectively utilize channel, temporal, and spatial information at different scales. Subsequently, the efficient temporal attention (ETA) module is introduced into the Fast path to enhance the model's detection performance and help the model better capture action information. Based on experimental results, the method with the addition of MSTA and ETA improved the mAP by 5.63% compared to the original SlowFast. The study [3] proposed a ConvNet model consisting of a spatial stream ConvNet using Inception V3 to extract spatial features and a temporal stream ConvNet built with

a CNN comprising five convolutional layers with 3×3 kernels, a stride of 1, and the ReLU activation function to extract temporal features. The extracted features were then concatenated to form a feature vector for the final prediction.

In [41], the authors proposed a novel methodology based on a Histogram of Actions combined with gaze information. A 3D-CNN was employed to extract spatiotemporal features from 2-minute video segments. The model achieved an F1-score of up to 90% on a dataset containing 1,414 clips across 13 action categories, demonstrating the effectiveness of integrating action frequency with students' attention direction to assess learning quality. In [42], the authors adopted the X3D architecture and introduced a fusion mechanism that combines RGB features with human skeleton data. This design enhances recognition performance under challenging conditions such as student occlusion and limited computational resources. Experiments conducted on a large-scale classroom behavior dataset showed that the proposed model achieved a Top-1 accuracy of 88.36%, while significantly reducing the number of parameters and computational cost compared to conventional 3D CNN models.

Some clip-level methods have been proposed for student activity recognition. However, these approaches typically rely on the assumption that the clips have be defined a prior. The main drawback of sequence-level methods is that in a classroom setting, where many students may be performing different activities at different times, determining the appropriate clips for analysis is not a straightforward task.

4.3. Continuous activity recognition approach

Methods in the continuous activity recognition category aim to detect complete instances of each activity in untrimmed classroom videos. This involves identifying the spatial location of each activity instance within the video frame, typically defined by a bounding box, as well as determining its start and end times. Figure 5 shows the input and output of continuous student activity recognition.

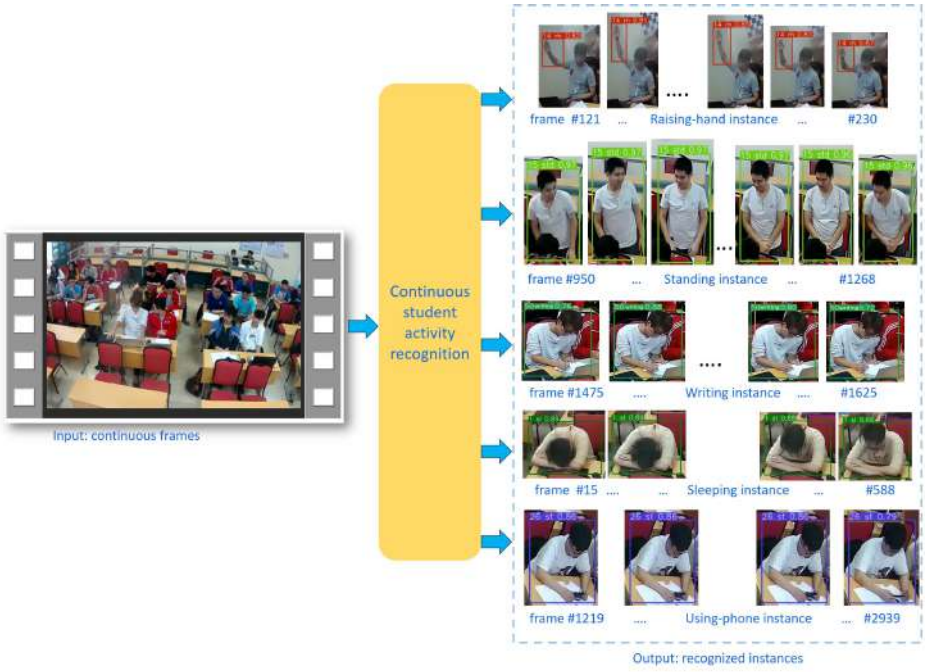

Figure 5. Continuous-level activity recognition approaches

This approach is similar to spatio-temporal action detection in action recognition [43]-[45], however due to the challenges of student activity recognition, very few works have been proposed for continuous recognition from untrimmed videos[24].

In the study [24], the author proposed a method for continuous hand-raising recognition from untrimmed video. The proposed approach combines hand-raising detection and tracking to follow the hand-raising activity over time. The detection model is an improved version of the Libra-RCNN deep learning object detection

network, incorporating an enhanced non-local block and leveraging correlation features based on distance and position between elements on the feature map. The tracking component utilizes the SORT object tracking module [46]. Table 2 lists all methods belonging to clip and continuous categories.

Table 2. Summary of methods belonging to clip-level and continuous approach for student activity recognition

Method	Description	Activities	Result	Recognition Level
Lei *et al.* (2019) [38]	Combine traditional features L2EMG (Local Log-Euclidean Multivariate Gaussian) and SIFT (Scale-Invariant Feature Transform)	Raising hand, standing, sitting, writing, reading	Achieved an average accuracy of 82.05%	Frame-level
Sharma *et al.* (2021) [39]	Use two-stream I3D-ResNet-50 model (RGB and Optical Flow) for video classification	20 activities (arguing, clapping, eating, hand raise, hitting, etc)	Accuracy of 72.3% on EduNet dataset	Clip-level
Nguyen *et al.* (2023) [23]	Combine Alphapose joint estimation with the DD-Net activity recognition network for skeleton-based activity recognition	Raising hand, sleeping, standing, sitting, using phone	Accuracy, precision, recall and F1-score were 74.58%, 78.71%, 74.58%, and 72.15% on StudentAct Skel	Clip-level
Zhang *et al.* (2023) [22]	Combine the Multi-scale Spatial-Temporal Attention (MSTA) and the Efficient Temporal Attention (ETA) to focus on salient features over time	7 activities	mAP of 91.1% on SCSB dataset	Clip-level
Liu *et al.* (2023) [3]	Propose a ConvNet model to extract spatial and temporal features consisting of a Spatial Stream ConvNet and a Temporal Stream ConvNet	7 activities	Overall accuracy of 83.01% on EduAction dataset	Clip-level
Ahmed *et al.* (2023) [41]	Introduce a method based on a Histogram of Actions combined with gaze data	13 activities (raising hand, yawn, writing, etc)	Achieved F1-score of up to 90%	Clip-level
Zou *et al.* (2025) [42]	Integrate a fusion mechanism between RGB features and human skeleton data to enhance recognition performance	6 activities (raising hand, standing up, listening, etc)	Achieved an Top-1 Accuracy of 88.36%	Clip-level
Le *et al.* (2023) [24]	Combine the improved Libra-RCNN model with the SORT tracking module to recognize and track hand-raising gestures	Raising hand	Frame-wise accuracy, Temporal IoU, F1-score, and Levenshtein score were 90%, 84.4%, 83.2% and 84.3%	Continuous
Nguyen *et al.* (2025) [47]	Propose a two-step method (STrack4Re) for continuous student activity recognition	5 activities	YOLOv5 and OC-SORT were selected for the object detection and tracking steps	Continuous
Bui *et al.* (2025) [48]	Enhanced the STrack4Re method by incorporating bounding boxes from virtual trajectories and integrating appearance features to increase tracking robustness.	5 activities	Achieved the highest F1-scores in activity recognition, outperforming the baseline by up to 10.9%.	Continuous

4.4. Classroom activity recognition image and video datasets

To train the recognition models and to evaluate their performance, several datasets have been collected and annotated. In the study by [29], the authors built a large-scale hand-raising dataset by recording videos from two static cameras in a real classroom with more than 30 students. After the annotation process, the dataset included a total of 40,000 hand-raising samples, with 28,000 samples used for training and 12,000 for testing. This dataset also presented challenges for the hand-raising detection task, such as occlusions, low resolution, and variations in hand-raising gestures. In [49], a dataset for hand-raising recognition at the frame level was introduced, comprising a total of 22,000 images with 76,000 bounding box annotations. All images had a resolution of 1920 $\times$ 1080 pixels. The training set was collected from four classes in an elementary school, while the testing set consisted of a video featuring 23 students in a meeting room. The significant differences between the postures of children and adults introduced certain challenges for hand-raising detection. Another similar dataset for hand-raising activity, named Class_HRP was also introduced in [24], consisting of 6,733 images recorded over multiple sessions in a 50-student classroom of the university. A total of 53,798 hand-raising samples have been annotated, with 4,928 images containing 34,511 samples used for the training set and 1,805 images with 19,287 samples allocated for the test set.

In [20], the authors constructed a large-scale classroom student behavior dataset collected from more than 30 primary and secondary schools. The dataset includes 70,000 hand-raising instances, 20,000 standing instances, and 3,000 sleeping instances, all annotated with bounding boxes. It has been regarded as one of the most challenging benchmarks due to substantial scale variations, severe class imbalance, and relatively low-quality annotations. The scale variation is evident in the significant difference between hand-raising samples (approximately 40×40 pixels) and standing samples (approximately 200×200 pixels). Moreover, about 70% of the objects occupy less than 0.5% of the total image area. These factors collectively make the dataset particularly challenging for frame-level student activity recognition. A dataset named StudentAct was introduced in [2]. The recording system utilized five cameras installed at different locations. The videos was collected in a 100 m² classroom with a ceiling height of 3.5 m and a capacity of 60 students, focusing on five primary activities, including standing (talking to the teacher), sitting, using phone, sleeping, and raising hand. StudentAct dataset comprises 596,371 bounding boxes annotated from 31,046 images. The dataset poses several challenges, including significant variations in bounding box sizes, severe occlusions due to high student density (20–30 per class), and class imbalance, where sitting has the highest number of samples (34,102), while standing has the least (3,259). These factors contribute to the complexity of student action recognition, making StudentAct a valuable benchmark for future research.

Several datasets have also been developed to support sequence-level student activity recognition. One such dataset, Beijing Normal University Large-scale classroom student action database (BNU-LCSAD), was introduced to facilitate the recognition, detection, and captioning of student behaviors in classroom settings [50]. The videos were collected from real classroom sessions at Beijing Normal University. In total, 4.5K video clips from different disciplines and different educational stages have been selected. EduNet, introduced in [39], comprises 7,851 manually annotated clips extracted from YouTube videos and recorded in an actual classroom environment, covering 20 action classes. Another dataset was introduced in [51]. This dataset contains 4,917 images covering 10 learner states: raising hand, standing up, clapping, taking photos, looking up, holding cheek, taking notes, playing with a mobile phone, stretching, and lying on the desk. In [3], the authors introduced a new dataset named EduAction, which focuses on college students' actions. The dataset includes seven action categories and 718 action clips collected in real classroom environments. More recently, [22] presented the SCSB dataset, which targets seven common student behaviors, including looking at the board, looking down, turning the head or body, talking, standing up, raising hands, and lying on the table. Approximately 250 minutes of classroom video were collected. They are then segmented into 10-second clips. Over 600 clips were annotated, each containing from 7 to 20 students, resulting in a total of 51,387 annotations across the dataset.

Recently, in a newest research [52], authors has presented a video dataset of student actions. The dataset includes 4,324 3-second clips depicting realistic and diverse actions during classes in various subjects at different educational levels, from kindergarten and elementary to middle school, with resolutions of 720p and 1080p. The dataset is categorized into 5 types of actions with 15 different labels. However, some labels are highly similar, posing significant challenges for recognition models. The authors in [53] have constructed a new multimodal dataset focusing on activity recognition in classroom surveillance images called activity recognition in classrooms (ARIC). The ARIC dataset has the advantages of having multiple perspectives, 32 activity types, three modalities, and real classroom scenarios. Table 3 summarizes the datasets available for student activity recognition.

Although several datasets have been collected to evaluate student activity recognition methods, they are all designed for frame-level or sequence-level recognition. To the best of our knowledge, no publicly available dataset supports continuous activity recognition. Furthermore, most existing datasets are private, which limits the ability to fairly compare different methods on a shared benchmark.

4.5. Open issues and future directions

Despite significant advancements in student activity recognition from videos, several open challenges remain. The first challenge is data scarcity and diversity, as existing datasets often lack sufficient variations in classroom settings, cultural contexts, and student behaviors. Additionally, data annotation is a time-consuming and labor-intensive process. As a result, despite the vast amount of video data collected over the years, the availability of annotated datasets remains limited. Future research should prioritize the collection and annotation of large-scale, diverse datasets to enhance model generalization. Moreover, establishing benchmark datasets is crucial to facilitate the comparison of recognition methods. In the broader field of activity recognition, datasets

such as NTU 60 and NTU 120 are widely used. However, in the domain of student activity recognition, there is still a lack of a large-scale, standardized benchmark dataset, which hinders the development and evaluation of more effective recognition models.

Secondly, the majority of current research focuses on frame-level detection, driven by advancements in object detection models. However, few studies have explored clip-based or continuous activity recognition. Frame-level predictions alone are often insufficient for meaningful classroom assessment, as they lack temporal coherence and context. Therefore, future efforts should prioritize continuous recognition approaches that can effectively capture and analyze student activities over extended time periods, enabling more accurate monitoring and evaluation.

Table 3. Summary of methods belonging to clip-level and continuous approach for student activity recognition

Dataset	#Activities	Statistical Information	Recognition Level	Availability
Si *et al.* (2019) [29]	1 (hand-raising)	40,000 samples (28,000 training, 12,000 testing)	Frame-level	Private
Liu *et al.* (2020) [49]	1 (hand-raising)	22,000 images with 76,000 bounding boxes, elementary school and university meeting room	Frame-level	Private
Class HRP (2023) [24]	1 (hand-raising)	6,733 images, 53,798 bounding boxes (34,511 training and 19,287 testing), university classroom	Frame-level	Private
Zheng Rui *et al.* (2020) [20]	3 activities (hand-raising, standing, sleeping)	70,000 hand-raising, 20,000 standing, 3000 sleeping boxes, primary and secondary schools	Frame-level	Private
StudentAct (2022) [2]	5 activities (standing, sitting, using phone, sleeping, and raising hand)	596,371 bounding boxes from 31,046 images, university classroom	Frame-level	Available upon request
Che *et al.* (2022) [51]	10 activities (e.g., raising hand, standing up, taking notes, etc.)	4,917 images, three angles and in day, night, fluorescent lamp lighting condition	Frame-level	Private
BNU-LCSAD (2021) [50]	11 activities (e.g., eating or drinking, listening, taking notes, using mobile phones, yawning, etc)	128 videos of 45min, real classroom scenes	Sequence-level	Private
EduNet (2021) [39]	20 activities (e.g., hand raiseing, hitting, etc.)	7,851 annotated clips collected from YouTube and actual classroom	Sequence-level	Available on request
EduAction (2023) [3]	7 activities (e.g., sleeping, listening to lectures)	718 action clips in college classroom environment	Sequence-level	Private
SCSB (2023) [22]	7 activities (e.g., looking at board, looking down)	600 labeled clips from 250 mins of videos with 51,387 annotations	Sequence-level	Private

Real-time recognition and efficiency remain significant challenges, particularly in resource-constrained environments where computational power is limited. Developing lightweight and efficient deep learning models is essential to enable real-time processing without compromising accuracy. Furthermore, most existing studies focus on activity recognition within a single classroom, limiting their scalability. Deploying a large-scale monitoring system across an entire school or university requires additional considerations, such as efficient allocation of computing resources, distributed processing, and network infrastructure optimization. Future research should explore scalable architectures and resource management methods to ensure the feasibility of large-scale deployment in educational settings.

Besides visual information, additional data sources such as audio, gaze tracking, and physiological signals can be collected from the classroom to provide a more cues for student activities understanding. A

promising future direction is multi-modal fusion, where integrating these complementary modalities alongside video can significantly enhance recognition accuracy.

Furthermore, context-awareness is crucial, as current models often fail to understand the broader classroom dynamics, such as interactions between students and teachers. Future work should explore graph-based or attention-driven models to incorporate contextual relationships.

Lastly, ensuring privacy and ethical considerations in student video analysis is essential, requiring privacy-preserving techniques such as federated learning. Addressing these challenges will contribute to more robust, scalable, and ethically responsible student activity recognition systems. .

5. CONCLUSIONS AND FUTURE WORKS

Student activity recognition from videos has obtained significant attention thanks to its potential applications in smart classrooms, adaptive learning, and educational analytics. This survey reviewed state-of-the-art approaches, categorizing them into frame-level, clip-level, and continuous recognition methods. Additionally, we provided an overview of available datasets for student activity recognition. The survey has highlighted that while promising progress has been made, several challenges remain, including data scarcity, real-time efficiency, multi-modal fusion, large-scale deployment and privacy. Finally, we have proposed suggestions for future research, aiming to advance the development of more robust and scalable student activity recognition systems.

ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4064.

REFERENCES

[1] S. Jahagirdar and R. Phalnikar, "Comparison of feed forward and cascade forward neural networks for human action recognition," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 25, no. 2, pp. 892-899, 2022, doi: 10.11591/ijeecs.v25.i2.pp892-899.

[2] P.-D. Nguyen *et al.*, "A new dataset and systematic evaluation of deep learning models for student activity recognition from classroom videos," in *2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2022, pp. 1–6, doi: 10.1109/MAPR56351.2022.9924673.

[3] K. Liu *et al.*, "EduAction: A college student action dataset for classroom attention estimation,"*Advanced Intelligent Computing Technology and Applications: 19th International Conference, ICIC 2023, Zhengzhou, China, August 10–13, 2023, Proceedings, Part IV*, 2023, pp. 237–248, doi: 10.1007/978-981-99-4752-2-20.

[4] L. Xiao, K. Luo, J. Liu, and A. Foroughi, "A hybrid deep approach to recognizing student activity and monitoring health physique based on accelerometer data from smartphones," *Scientific Reports*, vol. 14, no. 1, p. 14006, 2024, doi: 10.1038/s41598-024-63934-8.

[5] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010, [Online]. Available: ttps://doi.org/10.1016/j.imavis.2009.11.014.

[6] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017, [Online]. Available: https://doi.org/10.48550/arXiv.1605.04988.

[7] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image and Vision Computing*, vol. 55, pp. 42–52, 2016, [Online]. Available: https://doi.org/10.1016/j.imavis.2016.06.007.

[8] T. F. N. Bukht, H. Rahman, M. Shaheen, A. Algarni, N. A. Almujally, and A. Jalal, "A review of video-based human activity recognition: theory, methods and applications," *Multimedia Tools and Applications*, pp. 1–47, 2024, [Online]. Available: https://doi.org/10.1007/s11042-024-19711-w.

[9] N. A. Rahmad, M. A. As'ari, N. F. Ghazali, N. Shahar, and N. A. J. Sufri, "A survey of video based action recognition in sports," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, 2018, [Online]. Available: http://doi.org/10.11591/ijeecs.v11.i3.pp987-993.

[10] H. Yin, R. O. Sinnott, and G. T. Jayaputera, "A survey of video-based human action recognition in team sports," *Artificial Intelligence Review*, vol. 57, no. 11, p. 293, 2024, [Online]. Available: https://doi.org/10.1007/s10462-024-10934-9.

[11] Q. Liu, X. Jiang, and R. Jiang, "Classroom behavior recognition using computer vision: a systematic review," *Sensors*, vol. 25, no. 2, 2025, doi: 10.3390/s25020373.

[12] T. S. Nazaré and M. Ponti, "Hand-raising gesture detection with Lienhart-Maydt method in videoconference and distance learning," in *Iberoamerican Congress on Pattern Recognition*, 2013, pp. 512–519, [Online]. Available: https://doi.org/10.1007/978-3-642-41827-3-64.

[13] J. Jesna, A. S. Narayanan, and K. Bijlani, "Automatic hand raise detection by analyzing the edge structures," in *International Conference on Emerging Research in Computing, Information, Communication and Applications*, 2016, pp. 171–180, [Online]. Available: https://doi.org/10.1007/978-981-10-4741-1-16.

[14] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in

Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 761–769, [Online]. Available: https://doi.org/10.48550/arXiv.1604.03540.

[15] H. Zhou, F. Jiang, and R. Shen, "Who are raising their hands? hand-raiser seeking based on object detection and pose estimation," in *Asian Conference on Machine Learning*, 2018, pp. 470–485, [Online]. Available: https://api.semanticscholar.org/CorpusID:53611083.

[16] Z. Wang, F. Jiang, and R. Shen, "An effective yawn behavior detection method in classroom," in *International conference on neural information processing*, 2019, pp. 430–441, [Online]. Available: https://doi.org/10.1007/978-3-030-36708-4-35.

[17] W. Li, F. Jiang, and R. Shen, "Sleep gesture detection in classroom monitor system," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7640–7644, [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8683116.

[18] W. Liao, W. Xu, S. Kong, F. Ahmad, and W. Liu, "A two-stage method for hand-raising gesture recognition in classroom," in *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, 2019, pp. 38–44, [Online]. Available: https://doi.org/10.1145/3318396.3318437.

[19] R. Zheng, F. Jiang, and R. Shen, "Intelligent student behavior analysis system for real classrooms," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 9244–9248, doi: 10.1109/ICASSP40776.2020.9053457.

[20] R. Zheng, F. Jiang, and R. Shen, "GestureDet: Real-time student gesture analysis with multi-dimensional attention-based Detector.," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2020, pp. 680–686, [Online]. Available: https://doi.org/10.24963/ijcai.2020/95.

[21] F.-C. Lin, H.-H. Ngo, C.-R. Dow, K.-H. Lam, and H. L. Le, "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," *Sensors*, vol. 21, no. 16, p. 5314, 2021, [Online]. Available: https://doi.org/10.3390/s21165314.

[22] S. Zhang *et al.*, "MSTA-SlowFast: a student behavior detector for classroom environments," *Sensors*, vol. 23, no. 11, p. 5205, 2023, [Online]. Available: https://doi.org/10.3390/s23115205.

[23] P.-D. Nguyen *et al.*, "Skeleton-based student activities recognition from classroom videos," in *International Conference on Advances in Information and Communication Technology*, 2023, pp. 292–299, [Online]. Available: https://doi.org/10.1007/978-3-031-50818-9-32.

[24] T.-H. Le *et al.*, "Spatial and temporal hand-raising recognition from classroom videos using locality, relative position-aware non-local networks and hand tracking," *Vietnam Journal of Computer Science*, vol. 10, no. 02, pp. 243–271, 2023, [Online]. Available: https://doi.org/10.1142/S2196888822500397.

[25] J. Shin, N. Hassan, A. S. M. Miah, and S. Nishimura, "A comprehensive methodological survey of human activity recognition across diverse data modalities," *Sensors*, vol. 25, no. 13, 2025, doi: 10.3390/s25134028.

[26] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 379–387, [Online]. Available: https://doi.org/10.48550/arXiv.1605.06409.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proceedings of the 29th International Conference on Neural Information Processing Systems*, Vol. 1, 2015, pp. 91–99, [Online]. Available: https://doi.org/10.48550/arXiv.1506.01497.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[29] J. Si, J. Lin, F. Jiang, and R. Shen, "Hand-raising gesture detection in real classrooms using improved R-FCN," *Neurocomputing*, vol. 359, pp. 69–76, 2019, [Online]. Available: https://doi.org/10.1016/j.neucom.2019.05.031.

[30] T.-H. Le *et al.*, "Locality and relative distance-aware non-local networks for hand-raising detection in classroom video," in *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2021, pp. 1–6, doi: 10.1109/MAPR53640.2021.9585284.

[31] B. Bühler *et al.*, "Automated hand-raising detection in classroom videos: A view-invariant and occlusion-robust machine learning approach," in *International Conference on Artificial Intelligence in Education*, 2023, pp. 102–113, [Online]. Available: https://doi.org/10.1007/978-3-031-36272-9-9.

[32] J. Chen, M. Wang, L. Wang, and F. Huang, "Student motivation analysis based on raising-hand videos," *Sensors*, vol. 24, no. 14, p. 4632, 2024, [Online]. Available: https://doi.org/10.3390/s24144632.

[33] Q. T. Nguyen, H. T. Binh, T. D. Bui, and P. D. NT, "Student postures and gestures recognition system for adaptive learning improvement," in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, 2019, pp. 494–499, [Online]. Available: https://doi.org/10.1109/NICS48868.2019.9023896.

[34] A. Deshpande and V. Deshpande, "Student activity recognition in classroom environments using transfer learning," in *2023 International Conference on Computational Intelligence, Networks and Security (ICCINS)*, 2023, pp. 1–6.

[35] L. Li, M. Liu, L. Sun, Y. Li, and N. Li, "ET-YOLOv5s: Toward deep identification of students' in-class behaviors," *IEEE Access*, vol. 10, pp. 44200–44211, 2022, doi: 10.1109/ACCESS.2022.3169586.

[36] Q. Jia and J. He, "Student behavior recognition in classroom based on deep learning," *Applied Sciences*, vol. 14, no. 17, 2024, doi: 10.3390/app14177981.

[37] C. Wang and J. Yan, "A comprehensive survey of RGB-based and skeleton-based human action recognition," *IEEE Access*, vol. 11, pp. 53880–53898, 2023, doi: 10.1109/ACCESS.2023.3282311.

[38] F. Lei, Y. Wei, J. Hu, H. Yao, W. Deng, and Y. Lu, "Student action recognition based on multiple features," in *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2019, pp. 428–432, doi:

[39] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "EduNet: A new video dataset for understanding human activity in the classroom environment," *Sensors*, vol. 21, no. 17, p. 5699, 2021, [Online]. Available: https://doi.org/10.3390/s21175699.

[40] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733, doi: 10.1109/CVPR.2017.502.

[41] A. Abdelkawy, I. Alkabbany, A. Ali, and A. A. Farag, "Measuring student behavioral engagement using histogram of actions," *Pattern Recognit. Lett.*, vol. 186, pp. 337–344, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:274035283.
[42] S. Zou, D. Wu, J. Gan, J. Zhou, and J. Mei, "Lightweight classroom student action recognition method based on spatiotemporal multimodal feature fusion," *Computers, Materials & Continua*, vol. 83, no. 1, pp. 1101–1116, 2025, doi: 10.32604/cmc.2025.061376.
[43] P. Wang, F. Zeng, and Y. Qian, "A survey on deep learning-based spatio-temporal action detection," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 22, no. 04, p. 2350066, 2024, doi: 10.1142/S0219691323500662.
[44] O. Köpüklü, X. Wei, and G. Rigoll, "You only watch once: A unified CNN architecture for real-time spatiotemporal action localization," ArXiv, vol. abs/1911.0, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:208076597.
[45] J. Zhao *et al.*, "TubeR: Tubelet transformer for video action detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13588–13597, doi: 10.1109/CVPR52688.2022.01323.
[46] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*, 2016, pp. 3464–3468, [Online]. Available: https://doi.org/10.1109/ICIP.2016.7533003.
[47] P.-D. Nguyen, N. Le, K.-H. Bui, H.-Q. Nguyen, H.-Q. Nguyen, and T.-L. Le, "A method for continuous student activity recognition from classroom videos," *Machine Learning and Cybernetics*, (2025), vol. 16, pp. 7913–7937, 2025, [Online]. Available: https://doi.org/10.1007/s13042-025-02695-w.
[48] K.-H. Bui, P.-D. Nguyen, H.-H. Bui, Q.-T. Nguyen, and T.-L. Le, "Enhancing continuous student activity recognition through virtual trajectory and appearance matching," 2025, [Online]. Available: https://doi.org/10.1109/MAPR67746.2025.11133996.
[49] T. Liu, F. Jiang, and R. Shen, "Fast and accurate hand-raising gesture detection in classroom," *International Conference on Neural Information Processing*, pp. 232–239, 2020, [Online]. Available: https://doi.org/10.1007/978-3-030-63820-7-26.
[50] B. Sun *et al.*, "Student Class behavior dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes," *Neural Computing and Applications*, vol. 33, pp. 8335–8354, 2021, [Online]. Available: https://doi.org/10.1007/s00521-020-05587-y.
[51] B. Che, X. Li, Y. Sun, F. Yang, P. Liu, and W. Lu, "A database of students' spontaneous actions in the real classroom environment," *Computers and Electrical Engineering*, vol. 101, p. 108075, 2022, [Online]. Available: https://doi.org/10.1016/j.compeleceng.2022.108075.
[52] Z. Tan *et al.*, "Towards student actions in classroom scenes: New dataset and baseline," arXiv preprint arXiv:2409.00926, 2024, [Online]. Available: https://doi.org/10.48550/arXiv.2409.00926.
[53] L. Xu *et al.*, "ARIC: An activity recognition dataset in classroom surveillance images," arXiv preprint arXiv:2410.12337, 2024, [Online]. Available: https://doi.org/10.48550/arXiv.2410.12337.

BIOGRAPHIES OF AUTHORS

Phuong-Dung Nguyen graduated from Vietnam National University, Hanoi, with a bachelor's degree in information technology. She received her master's degree in information systems from Hanoi National University of Education. Currently, she is working at Thuyloi University while pursuing a PhD in Hanoi University of Science and Technology. Her research interests include image processing, computer vision, and video understanding. You can contact her via email: dungntp@tlu.edu.vn.

Khanh-Huyen Bui is currently final-year student majoring in Electronics and Telecommunications Engineering at School of Electrical and Electronic Engineering (SEEE), Hanoi University of Science and Technology (HUST). Her research interests include image processing, computer vision, and video understanding. You can contact her via email: huyen.bk210455@sis.hust.edu.vn.

Thi-Lan Le graduated in Information Technology from Hanoi University of Science and Technology (HUST), Vietnam. She obtained MS. degree in Signal Processing and Communication from HUST, Vietnam. In 2009, she received her Ph.D. degree at INRIA Sophia Antipolis, France in video retrieval. She is currently associate professor at School of Electrical and Electronic Engineering (SEEE), HUST, Vietnam. Her research interests include images processing, computer vision, content-based indexing and, retrieval, video understanding and human-robot interaction. She can be contacted at email: lan.lethi1@hust.edu.vn.