JST: Smart Systems and Devices
Volume 33, Issue 3, September 2023, 016-023

BK-SAD: A Large Scale Dataset for Student Activity Recognition

Nguyen Van Giang*, Nguyen Chan Hung, Nguyen Quang Dich, Trinh Cong Dong
Institute of Control Engineering and Automation, Hanoi University of Science and Technology, Ha Noi, Vietnam
**Corresponding author email: giangnv.gnvml@outlook.com*

Abstract

Skeleton-based human action recognition has emerged as a prominent research topic in the field of artificial intelligence due to its broad applicability in a wide range of domains, including but not limited to healthcare, security and surveillance, entertainment, and intelligent environments. In this paper, we propose a novel data collection methodology and present BK-Student Activity Dataset (BK-SAD), a new 2D dataset for student activity recognition in smart classrooms that outperforms the existing NTU RGB+D 120 dataset, SBU Kinect Interaction dataset. Our dataset contains three classes: hand raising, doze-off, and normal activities. The dataset was collected using cameras placed in real classroom environments and consisted of video data from multiple viewpoints. The dataset contains over 2700 videos of students raising their hands, over 1700 videos of students dozing off during class, and over 8500 videos of normal activities. In addition, to evaluate the effectiveness of the proposed dataset, we give some baseline performance figures for neural network architectures trained and tested for student activity recognition on BK-SAD dataset. These ConvNet architectures demonstrate significant performance improvement on the proposed dataset. The effectiveness of the proposed novel data collection methodology and BK-SAD dataset in this paper will enable further research and development of activity recognition models for classroom environments, with potential applications in smart education and intelligent classroom management systems.

Keywords: Dataset, action recognition, skeleton pose, student activity recognition, smart classroom.

1. Introduction

Human activity recognition (HAR) has emerged as a crucial task in the field of artificial intelligence (AI). HAR systems aim to identify and classify human actions based on sensor data, such as video or sensor readings. HAR has a wide range of applications, spanning from healthcare to security and entertainment [1, 2]. In recent years, there has been growing interest in HAR systems for monitoring and managing student activity in educational environment.

Recognizing student actions, such as hand raising [3] and dozing off [4], is a critical task for the development of intelligent classroom management systems. These systems can help educators track student engagement, intervene when necessary, and improve overall educational outcomes. However, recognizing these subtle movements presents significant challenges in developing effective HAR systems for classrooms.

One of the primary challenges is the lack of large-scale labeled datasets for HAR in a school environment. While several benchmark datasets for HAR exist such as: NTU RGB+D 120, SBU Kinect Interaction dataset, they are often limited in terms of the number of action classes, camera views, and the diversity of the subjects. Moreover, the datasets often do not focus on the specific actions relevant to educational environments, such as hand raising and dozing off.

To address these challenges, this paper presents BK-SAD dataset for activity recognition in a school environment, focusing on the recognition of student hand raising, dozing off, and normal activities. The proposed dataset contains a large number of labeled video samples collected from multiple camera views and diverse student populations. BK-SAD is available at https://visedu.vn/en/bk-sad-dataset.

The contributions of this paper lie in the development of a novel data collection methodology as well as a high-quality BK-SAD dataset and the demonstration of the effectiveness of the proposed approach for improving student activity monitoring in classrooms. The remainder of this paper is structured as follows. Section 2 provides a review of the related work in HAR and the challenges in developing effective HAR systems for educational environments. Section 3 presents the details of the proposed dataset, including the data collection process, annotation, and statistics. Section 4 gives the performance of a number of ConvNet architectures that are trained and tested on the BK-SAD dataset, followed by the conclusion and future directions in Section 5.

ISSN: 2734-9373
https://doi.org/10.51316/jst.168.ssad.2023.33.3.3
Received: March 14, 2023; accepted: July 11, 2023

2. Related Work

HAR has been an active area of research for many years, with a wide range of applications and approaches. Early HAR systems typically used handcrafted features, such as histogram of oriented gradients (HOG) [5], to extract features from video or sensor data, which were then fed into a classifier for activity recognition. However, these methods often suffer from limited performance and scalability, as they do not capture the complex spatiotemporal relationships in human actions.

In recent years, deep learning-based approaches have shown notable performance in HAR, particularly for skeleton-based representations [6, 7]. Skeleton-based representations capture the spatial relationships between different body joints and their temporal evolution, providing a rich representation for activity recognition. Several deep learning models have been proposed for skeleton-based HAR, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and graph convolutional networks (GCNs).

Despite the recent advances in HAR, developing effective HAR systems for educational environments presents significant challenges. One of the primary challenges is the lack of labeled datasets for recognizing student actions in classrooms. Existing datasets NTU RGB+D dataset, NTU RGB+D 120 dataset, SBU Kinect Interaction dataset, often focus on more general activity recognition tasks, such as recognizing actions in sports or public spaces, and do not capture the specific actions and scenarios relevant to educational environments. Additionally, recognizing subtle movements, such as hand raising, doze-off, requires precise annotation and high-quality data, which can be difficult to collect in real-world classrooms.

To address the challenges in developing effective HAR systems for educational environments, we propose a novel data collection methodology and BK-SAD dataset for activity recognition in a school environment. The dataset focuses on recognizing student hand raising, dozing off, and normal activities, which are critical actions for student monitoring.

3. The Dataset

3.1. Data Collection

In the classroom environment, student behaviors can be broadly classified into three categories: positive learning, negative learning, and normal behavior. Among these, we have carefully chosen hand raising, dozing, and normal activities as discernible behavioral indicators representing positive, negative, and normal learning activities, respectively. However, certain behaviors like phone usage, socializing with peers, or diligent studying, while associated with either negative or positive learning behaviors, lack distinctive characteristics. Consequently, we have not obtained dedicated datasets for these particular behaviors.

The data collection process for activity recognition in a smart classroom environment is critical to ensuring the quality and representativeness of the dataset used to train and evaluate machine learning models. In this study, an extensive experiment was conducted to collect hundreds of classroom videos at Chu Van An Secondary School in Long Bien District, Vietnam.

The process of collecting data for activity recognition involved capturing videos of classroom activities using multiple cameras placed at strategic locations around the room. The cameras were positioned to provide a comprehensive view of each student's actions and captured high-resolution video footage that was later processed to extract key features for activity recognition.

To extract these features, the BK-Pose model (submitted to Journal of Science & Technique, LQDTU-JST) was used to analyze the video footage and extract skeleton data for each student in the classroom. The BK-Pose model is a lightweight and efficient approach to extracting key features from videos and is effective in identifying human poses.

To ensure the representativeness of the dataset, a diverse range of students was captured in the classroom videos, including students of different ages and genders. This was important to ensure that the resulting dataset would be able to generalize well to a range of different classroom environments and student populations.

Throughout the data collection process, careful attention was paid to ethical considerations, such as obtaining consent from the students and their parents, and ensuring that the data was anonymized to protect the privacy of the individuals captured in the videos. These ethical considerations are critical to ensuring that the resulting dataset can be used for research purposes while also respecting the rights and privacy of the individuals captured in the videos.

Table 1. Number of videos in each type of activity dataset

Class Label	Train videos	Test videos	% of Total Dataset
Doze	1217	522	13.32%
Hand-raising	1996	727	20.85%
Normal	6018	2579	65.83%
Total	9231	3828	100%

3.2. Data Collection Methodology

In this section, we describe the collection process how candidate videos were obtained from smart classrooms, and then the processing pipeline that was used to select the candidates and clean up the dataset.

3.2.1. Environment setup

The collection of activity data in smart classroom environments poses significant challenges, necessitating careful consideration and planning. In this study, we obtained the necessary permissions and consents from the administration of Chu Van An Secondary School to collect activity data from students across various grade levels. A total of 10 classes participated in the study, with each class comprising a minimum of 30 students. We collected data from each class within a fixed time frame of 20 minutes.

To ensure comprehensive coverage of student activities, we deployed a smart classroom with desks arranged in rounded configurations, facilitating easy interaction and communication among students. Additionally, we installed a minimum of four high-resolution 3MP cameras at strategic locations, including the top, left, right, and back corners of the classroom (Fig. 1, 2, 3, 4), to capture all corners of the classroom. By adopting this setup, we aimed to provide a consistent and controlled environment for collecting high-quality activity data from smart classrooms, thereby advancing the field of activity recognition in educational settings.

3.2.2. Videos collection

The collection of high-quality data is critical for the development and training of machine learning models that recognize human activities in smart classroom environments. In this study, we adopted a systematic approach to collecting diverse and representative data to ensure the generalizability and accuracy of our models.

To capture a range of student activities, we focused on three key actions: hand raising, doze-off, and normal. For each action, we conducted 100 repetitions, with each recorded video spanning a duration of 5 seconds. By incorporating multiple repetitions, we aimed to create a comprehensive dataset that accurately reflects the complex and dynamic nature of student activity.

To minimize biases and ensure the diversity of our dataset, we employed several strategies during data collection. Firstly, we systematically swapped students across different positions to prevent the overrepresentation of certain individuals or groups. Additionally, we randomized the order of the target activities across different positions, thereby reducing the potential for confounding effects. These measures aimed to capture the full range of student activity while reducing bias in our dataset.

Fig. 1. The front direction of the classroom.

Fig. 3. The right direction of the classroom.

Fig. 2. The left direction of the classroom.

Fig. 4. The back direction of the classroom.

Furthermore, we collected videos of students' normal activity during the data collection period. These videos were extracted concurrently with the recordings of the target actions, thus ensuring the consistency of our data collection. By incorporating normal activity videos, we aimed to create a more representative dataset that reflects the natural and varied activity of students in smart classroom environments.

3.2.3. Manual labeling process

Upon completion of video data collection, manual labeling of actions is necessary to generate a labeled dataset for training machine learning models. However, this process can be time-consuming and labor-intensive. To alleviate this issue, we utilized the Computer Vision Annotation Tool (CVAT) to automate and manage the labeling task (Fig. 5, 6, 7). This tool assisted in identifying the frame ranges that correspond to specific activities and labeling bounding boxes (bbox) to identify students in the video frames. The long videos were segmented into 10-second clips, each comprising 250 frames, resulting in 1200 labeling tasks for 10 classrooms. Each task required approximately a half day to complete with one person. Ten individuals worked continuously for two months to complete the labeling process.

For each frame, a bbox was defined to isolate each student and assign an activity label to that frame. This process allowed us to generate a comprehensive dataset that accounts for variations in student positions and activity. The manual labeling process ensured the accuracy and reliability of the dataset, resulting in a high-quality labeled dataset for machine learning purposes.

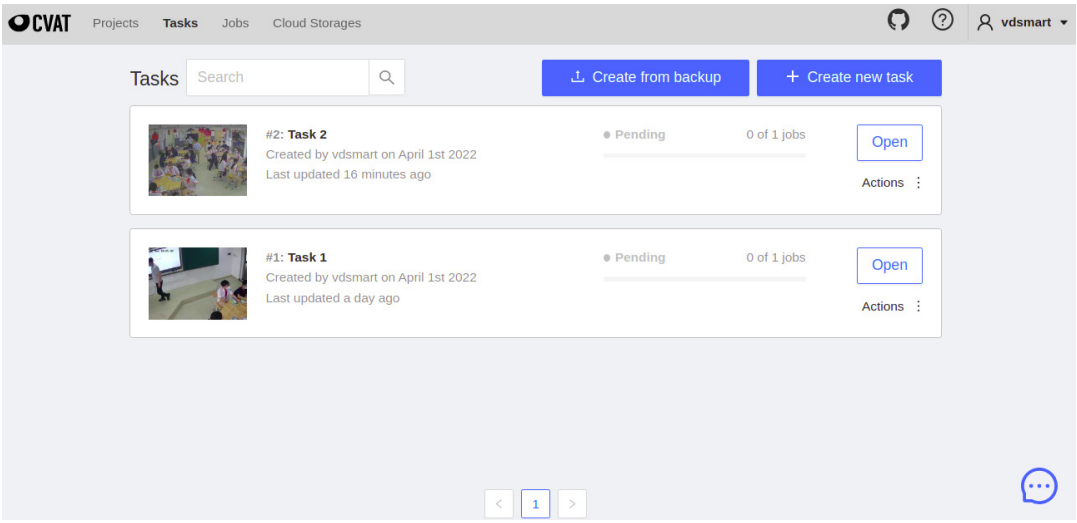

Fig. 5. CVAT's user interface.

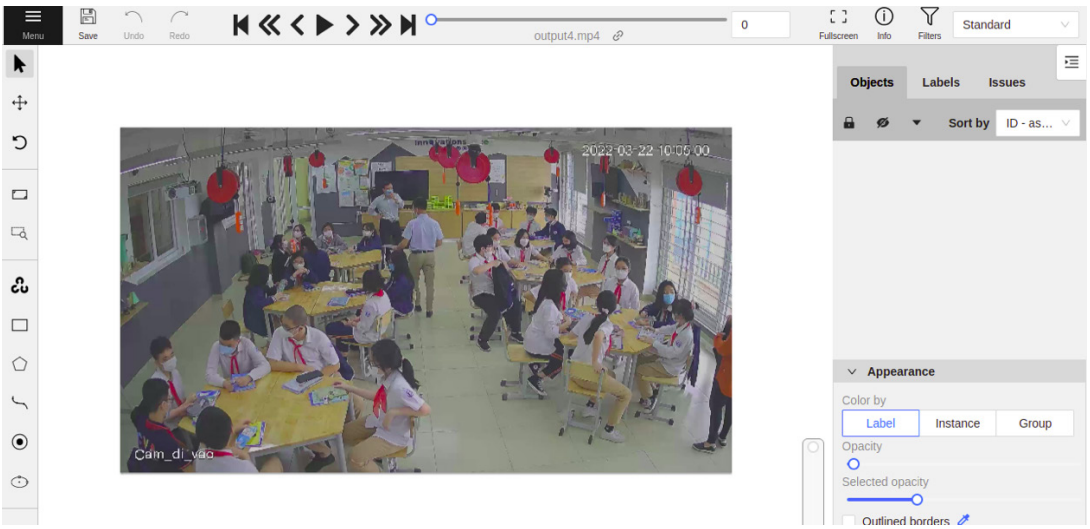

Fig. 6. An CVAT sample task in Chu Van An Secondary School

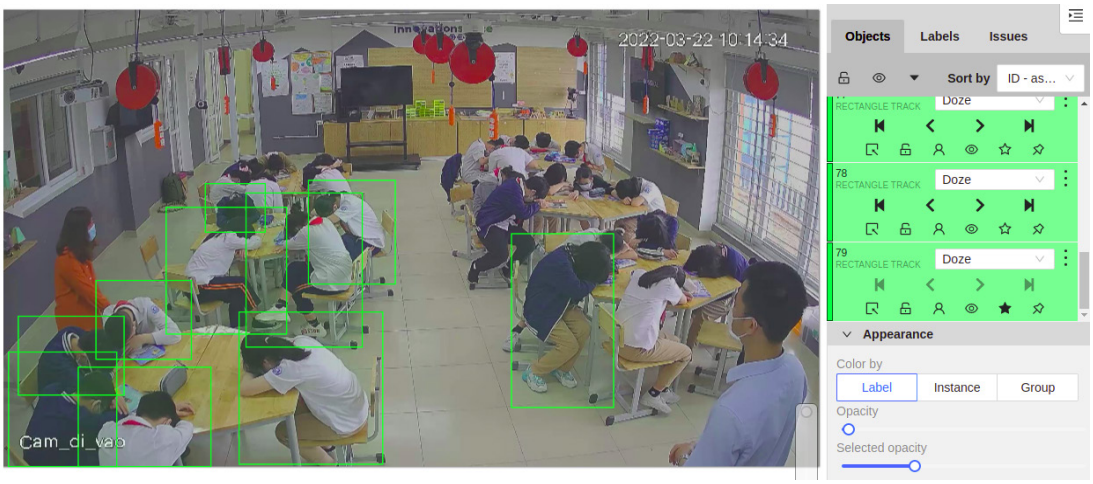

Fig. 7. Assigning activities to each student by manually labeling them.

To validate the labeling accuracy, we randomly selected a subset of labeled data and evaluated the labeling quality using the inter-annotator agreement metric. Our results indicated a high degree of labeling consistency, with an inter-annotator agreement score of 0.92, indicating a reliable labeling process.

3.2.4. Keypoint extraction

In this research study, we present a novel data collection methodology that utilizes the BK-Pose model to extract keypoint coordinates for each student in the smart classroom environment. The BK-Pose model is a lightweight neural network architecture trained on the Common Objects in Context (COCO) dataset, which has demonstrated exceptional accuracy in human pose estimation tasks. Our proposed methodology aims to automate the keypoint labeling process, significantly reducing the time and effort required for manual labeling. In contrast to manual labeling, our methodology employs advanced computer vision techniques to extract keypoint coordinates from bounding box picture frames. This technique involves the application of the BK-Pose model to extract 2D coordinates for 18 keypoints of each student in the image.

Our methodology begins by manually defining the bounding box around each student in the picture frame. The bounding box defines the region of interest that contains the student's body and allows us to extract the keypoint coordinates accurately. We then apply the BK-Pose model to extract the keypoint coordinates for each student in the image. The BK-Pose model utilizes a deep neural network architecture that can accurately estimate human poses by identifying the location of each body joint. By leveraging the BK-Pose model, our methodology can automatically extract keypoint coordinates for each student in the image, enabling us to analyze their body posture, movements, and behavior.

The BK-Pose model is an excellent choice for keypoint extraction due to its lightweight design and high accuracy. Compared to other models, such as the CMU-Pose [8] model, the BK-Pose model is approximately 55 times faster (compared to the FPS metric), enabling the rapid processing of a large number of frames. This is crucial in smart classroom environments where there are numerous students and data needs to be processed in real-time.

Our proposed methodology has several advantages over traditional manual labeling approaches. Firstly, it significantly reduces the time and effort required for data collection, enabling more efficient data collection and processing. Secondly, our methodology is less prone to errors, ensuring the accuracy and consistency of the data collected. Finally, by leveraging the BK-Pose model, our methodology can provide a more detailed and comprehensive understanding of student behavior and engagement levels.

4. Benchmark Performance

In this section, we first briefly describe two standard ConvNet architectures for human action recognition in video. We then use these architectures as baselines and compare their performance by training and testing on the BK-SAD dataset. We also include their performance on RGB+D 120 dataset, SBU Kinect Interaction dataset.

The BK-SAD dataset is novel and there is no existing dataset specifically related to capturing student activities in a smart classroom. We have also surveyed and explored relevant datasets concerning activity recognition, and we have found two datasets, namely SBU Kinect Interaction and NTU RGB+D 120, that have some activities that exhibit similarities to raising hands, doze-off, and normal activities. Therefore, we have selected these datasets to perform a comparative validation with the BK-SAD dataset.

Table 2. Baseline comparisons across datasets on three classes: hand-raising, doze off, normal

Method	SBU Kinect Interaction	NTU RGB+D 120	BK-SAD
ConvNet+LSTM	92.5	93.2	**93.7**
Two-Stream	93.7	94.1	**94.3**

Table 3. Confusion matrix of Two-Stream networks on BK-SAD dataset

Actual/Predicted	Hand-raising	Doze off	Normal
Hand-raising	13259	899	379
Doze	1187	8886	356
Normal	1074	465	50043

Table 4. Classification accuracy of Two-Stream networks on BK-SAD dataset.

Type of Dataset	Hand-raising	Doze off	Normal
Accuracy (%)	91.5	85.3	97.0

We consider two typical approaches for video classification: ConvNets with an LSTM on top [9, 10] and two-stream networks [11, 12]. There have been many improvements over these basic architectures, but our intention here is not to perform a thorough study on what is the very best architecture on the BK-SAD dataset, but instead to provide an indication of the level of difficulty of the dataset.

4.1. ConvNet+LSTM

The utilization of image classification networks for video analysis has gained significant attention due to their high performance levels. This has led to attempts to reuse them with minimal modifications by extracting features independently from each frame and pooling their predictions over the entire video, akin to bag-of-words image modeling approaches. However, this methodology neglects the temporal structure present in videos, leading to a lack of discriminative power in recognizing actions with temporal dynamics.

To address this limitation, a more comprehensive approach involves incorporating a recurrent layer, such as an LSTM, into the model to capture temporal ordering and long-range dependencies. To achieve this, we add an LSTM layer with batch normalization after the last average pooling layer of a ResNet-50 model, with 512 hidden units. A fully connected layer is then added to the output of the LSTM for multi-way classification, and at test time, the model output for the last frame is used for classification.

4.2. Two-Stream Networks

While LSTMs applied to features extracted from the final layers of ConvNets can effectively model high-level variation, they may not capture fine-grained, low-level motion, which is crucial in many video analysis tasks. Additionally, training such models can be computationally expensive due to the need to unroll the network through multiple frames for backpropagation-through-time.

A practical alternative approach was proposed by Simonyan and Zisserman, which involves modeling short temporal snapshots of videos by averaging predictions from a single RGB frame and a stack of 10 externally computed optical flow frames. This is achieved by passing the frames through two replicas of an ImageNet-pretrained ConvNet. The flow stream incorporates an adapted input convolutional layer with twice as many input channels as flow frames (since flow contains two channels, horizontal and vertical). During test time, multiple snapshots are sampled from the video, and the action prediction is obtained by averaging the predictions from the sampled snapshots. This approach has been demonstrated to achieve high performance on existing benchmarks while being efficient in both training and testing.

4.3. Implementation Details

In the ConvNet+LSTM and Two-Stream architectures, ResNet-50 serves as the base architecture. For the Two-Stream architecture, a separate ResNet-50 is trained independently for each stream. As previously mentioned, in our experiments with RGB+D 120 and SBU Kinect Interaction datasets, the ResNet-50 model is pre-trained on ImageNet, while in experiments with BK-SAD dataset, it is trained from scratch. In all cases, we utilized standard stochastic gradient descent (SGD) with momentum for training the models on videos, with synchronous parallelization across 2 GPUs A30 for all models. We trained the models on BK-SAD for up to 100,000 steps, with a 10x reduction of learning

rate when the validation loss reached saturation. We fine-tuned the weight decay and learning rate hyperparameters on the validation set of BK-SAD. We implemented all the models in TensorFlow.

4.4. Baseline Evaluations

This section presents a performance comparison of the two baseline architectures under different training and testing datasets.

The experimental results show that both ConvNet+LSTM and Two-Stream models achieve high accuracies across all three datasets. Among the three datasets, the BK-SAD dataset shows the highest accuracy for both models, with a score of 93.7% for ConvNet+LSTM and 94.3% for Two-Stream. When comparing the performance of ConvNet+LSTM and Two-Stream models on each dataset, we observe that the Two-Stream model consistently outperforms the ConvNet+LSTM model. Specifically, the Two-Stream model achieves 93.7%, 94.1%, and 94.3% accuracy on SBU Kinect Interaction, NTU RGB+D 120, and BK-SAD datasets, respectively, whereas the ConvNet+LSTM model achieves 92.5%, 93.2%, and 93.7% accuracy on the same datasets.

As far as class difficulty, the prevalence of occlusion in a classroom environment is a well-documented challenge in the field of computer vision, and it is particularly problematic in the context of pose estimation tasks. Our analysis highlights the need for developing more robust algorithms that can handle partial or complete occlusion, as this is critical for achieving higher accuracy in predicting body poses in a classroom environment. Our study contributes to the existing body of research on pose estimation in educational contexts, emphasizing the importance of addressing the challenges posed by occlusion in real-world settings. These findings have implications for the development of more accurate and reliable machine learning algorithms for analyzing classroom interactions and dynamics, thereby advancing the field of smart education.

In this study, we present a comprehensive experimental investigation aimed at obtaining a large corpus of classroom videos from Chu Van An Secondary School located in Long Bien District, Vietnam. Our methodology involves a novel data collection approach, which has proven to be highly effective in generating high-quality videos in a classroom environment.

The BK-SAD dataset is found to be well-suited for a variety of machine learning applications in the field of education. Our analysis of the collected data demonstrates the exceptional performance of our proposed data collection methodology and dataset in capturing student activity in smart classroom.

Furthermore, our findings suggest that the BK-SAD dataset is an invaluable resource for researchers and practitioners interested in developing and evaluating machine learning algorithms that target educational environments. Overall, the results of our experiment validate the efficacy and potential of our data collection approach and dataset for advancing the field of smart education.

Fig. 8. Student activity recognition at Chu Van An Secondary School, Label: Normal, Hand-raising, Doze from left to right respectively.

5. Conclusion

This paper presents a novel data collection methodology and the BK-SAD dataset, a 2D dataset for skeleton-based human action recognition (HAR) in smart classrooms. HAR has emerged as a prominent research topic in the field of artificial intelligence due to its broad applicability in various domains. The proposed dataset is particularly valuable for research and development of activity recognition models for classroom environments, with potential applications in smart education and intelligent classroom management systems.

We evaluated the effectiveness of the proposed dataset by performing baseline experiments using Convolutional Neural Network (ConvNet) architectures for student activity recognition. The

proposed dataset outperformed existing datasets such as the NTU RGB+D 120 and SBU Kinect Interaction datasets, demonstrating its value in improving the performance of activity recognition models in classroom environments.

Acknowlegement

We express our gratitude to the Ministry of Science and Technology for funding the national research project titled “Development of Digital Transformation model for Smart School" codename KC-4.0-06/19-25.

References

[1] Dhiman, C., Saxena, M., Vishwakarma, D. K., Skeleton-based view invariant deep features for human activity recognition, In Proceedings of the Fifth IEEE International Conference on Multimedia Big Data, Singapore, 11-13 September 2019, pp. 225-230. https://doi.org/10.1109/BigMM.2019.00-21

[2] Jiang, X., Xu, K., Sun, T., Action recognition scheme based on skeleton representation with DS-LSTM network, IEEE Trans. CircuitsSyst. Video Technol. 2020,30, 2129-2140. https://doi.org/10.1109/TCSVT.2019.2914137

[3] Jesna, J., Narayanan, A. S., Bijlani, K., Automatic hand raise detection by analyzing the edge structures. In Proceedings of the 4th International Conference on Emerging Research in Computing, Information, Communication and Applications, Bangalore, India, 29-30 July 2016, pp. 171-180. https://doi.org/10.1007/978-981-10-4741-1_16

[4] Li, W., Jiang, F., Shen, R, Sleep gesture detection in classroom monitor system. In Proceedings of the IEEE InternationalConference on Acoustics, Speech and Signal Processing (ICASSP 2019), Brighton, UK, 12-17 May 2019, pp. 7640-7644. https://doi.org/10.1109/ICASSP.2019.8683116

[5] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[6] Althloothi, S., Mahoor, M.H., Zhang, X., Voyles, R.M, Human activity recognition using multi-features and multiple kernellearning, Pattern Recognit. 2014, 47, 1800-1812. https://doi.org/10.1016/j.patcog.2013.11.032

[7] Cippitelli, E., Gasparrini, S., Gambi, E., Spinsante, S., A human activity recognition system using skeleton data from RGBDsensors, Comput. Intell. Neurosci. 2016, 4351435. https://doi.org/10.1155/2016/4351435

[8] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis amp, Machine Intelligence, vol. 43, no. 01, pp. 172-186, jan 2021. https://doi.org/10.1109/TPAMI.2019.2929257

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2625-2634, 2015. https://doi.org/10.1109/CVPR.2015.7298878

[10] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4694-4702, 2015. https://doi.org/10.1109/CVPR.2015.7299101

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman, Convolutional two-stream network fusion for video action recognition. In IEEE International Conference on Computer Vision and Pattern Recognition CVPR, 2016. https://doi.org/10.1109/CVPR.2016.213

[12] K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems, pages 568-576, 2014.