Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09741-5

Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus

M. Rashmi[1] · T. S. Ashwin[1] · Ram Mohana Reddy Guddeti[1]

Received: 5 August 2019 / Revised: 26 July 2020 / Accepted: 26 August 2020 /
Published online: 17 September 2020

Abstract
In the era of smart campus, unobtrusive methods for students' monitoring is a challenging task. The monitoring system must have the ability to recognize and detect the actions performed by the students. Recently many deep neural network based approaches have been proposed to automate Human Action Recognition (HAR) in different domains, but these are not explored in learning environments. HAR can be used in classrooms, laboratories, and libraries to make the teaching-learning process more effective. To make the learning process more effective in computer laboratories, in this study, we proposed a system for recognition and localization of student actions from still images extracted from (Closed Circuit Television) CCTV videos. The proposed method uses (You Only Look Once) YOLOv3, state-of-the-art real-time object detection technology, for localization, recognition of students' actions. Further, the image template matching method is used to decrease the number of image frames and thus processing the video quickly. As actions performed by the humans are domain specific and since no standard dataset is available for students' action recognition in smart computer laboratories, thus we created the STUDENT ACTION dataset using the image frames obtained from the CCTV cameras placed in the computer laboratory of a university campus. The proposed method recognizes various actions performed by students in different locations within an image frame. It shows excellent performance in identifying the actions with more samples compared to actions with fewer samples.

Keywords Human action recognition · Smart campus · Object detection · Object localization · Neural networks · Computer enabled laboratories

✉ M. Rashmi
nm.rashmi@gmail.com

T. S. Ashwin
ashwindixit9@gmail.com

Ram Mohana Reddy Guddeti
profgrmreddy@nitk.edu.in

[1] Department of Information Technology, National Institute of Technology Karnataka Surathkal, Mangalore, 575025, India

1 Introduction

Incorporating the human ability to recognize another person's actions to a machine is one of the challenging scientific research area in computer vision and machine learning. Automated video analysis systems can detect events related to human actions or human behavior and thus play an essential role in surveillance systems. Many researchers have worked on image and video analysis for the past several years to tackle different problems like object detection and localization, action recognition, and event recognition. HAR is a process in which the model interprets the action performed by a human, such as eating, smiling, cycling, etc. Automated video analysis systems that can interpret the events related to human actions play a major role in different domains. Several works are available in the existing literature in the area of HAR, which plays a vital role in real-world applications such as video surveillance [16, 37], video retrieval [40] and so on. HAR is a challenging task in computer vision due to the quality of the video, the direction from where the video captured, and the difference in speed of execution of the action by different people. Most of the existing works on HAR tried to deal with the aforementioned challenges by mining both spatial and temporal information from the video. The HAR from the videos can be categorized based on the way the video is analyzed. Some methods use the sequence of consecutive frames for action detection, such as [14], but many action categories can be explored from a single image also. Recently, some researchers have explored the actions from the still images [20, 31, 52].

1.1 Smart campus

"Smart Campus" is defined as integrating all kinds of application services, setting up wise, intelligent teaching, learning, and living environment to make it suitable for: teaching, scientific research, management, as well as forming an integrated system with the cooperation and self-adjustment capability, based on the Internet of Things (IoT) [19]. Due to advances in IoT technology, deep learning, and machine learning, the home, building, university campus, and cities are becoming smart. The smart campus design and implementation depend on campus needs. Some features of the smart campus are adapted from the smart city such as a smart room, a system that provides details about the classroom that is being used or not [44]. In a smart campus, campus staff can focus on their core tasks while most of the other operations are managed seamlessly to offer dynamic reports, that help decision making. A large number of researchers are working towards making human life easy in smart campus and smart cities.

Various approaches based on image and video analysis were proposed to interpret the student's emotion, engagement, learning attention, and so on. For example, facial emotion recognition, based on the Viola-Jones algorithm in the learning environment, is suggested in [6]. A reliable facial expression database is created in the online learning environments to meet the needs of automatic academic emotion inference [3]. A perspective-n-point method is developed to estimate the student's head pose for a single-image to achieve visualization of learning attention [20]. In [49], the authors suggested approaches for automated recognition of student's engagement based on facial expression. A mind wandering detector from video based on facial features for classroom and laboratory is developed in [4]. Student independent automated engagement detection based on features extracted from the video, heart rate, and animation units is developed in [34].

In the current scenario, the number of students in universities is increasing rapidly. It is essential to monitor students' activity to improve the teaching-learning environment within the campus. In the traditional method to monitor the students' activities in the campus,

human resources are used. A large number of surveillance cameras deployed in indoor as well as outdoor environments of smart campus for a different purpose. These cameras can be used for security, management, planning as well as to monitor students' activities inside the campus. These cameras can generate a huge amount of video data with different levels of complexity. So we need a system that can automate the process of the students' monitoring. These students' monitoring systems must be capable of recognizing and localizing students' activities inside the campus.

In the existing literature, several methods based on video/image analysis were proposed for different purposes, but none of these focus on recognizing students' activities inside the computer laboratory. To ensure the best utilization of resources in the computer laboratory, we need a system to monitor students' activities inside the laboratory.

To the best of our knowledge, we are the first to propose a system for monitoring students' activities inside a computer laboratory of a smart campus in Indian scenario. The main contributions of our proposed work are:

- To recognize the student actions with localization in real-time using YOLOv3[42] object localization and classification techniques.
- Multiple students' action recognition and localization in a single image frame.
- Create a dataset with five different students' actions for multiple students in a single image frame obtained from the CCTV cameras present in the computer science laboratories.
- Analyzing video obtained from CCTV cameras deployed in computer laboratories for easy monitoring of students as quickly and as accurately as possible.

The remaining part of the paper is organized as follows: Section 2 provides an overview of related work on the proposed model. Section 3 presents the details of the new dataset created for the proposed work. Section 4 gives the details of the proposed solution for student action recognition. Section 5 elaborates on the experiments conducted, and the corresponding results, and finally, Section 6 concludes the work with future direction.

2 Related work

As the scope of the smart campus is very broad, several methods were proposed in different areas of smart campus. Similarly, many solutions were recommended for a smart classroom. For example, in [29], authors specified technological and operational components of a future emotionally-aware Artificial Intelligent smart classroom that delivers automated real-time feedback to the teacher for improving the effectiveness of the presentation. An agent is designed to provide services of augmented reality to display and design other augmented scenarios in a smart classroom is proposed in [10]. A light-weight classroom scheduling architectural framework to reduce the burden of educational programs is proposed in [48]. An automated attendance management system based on face detection and recognition algorithms to detect student entering the classroom is proposed in [13].

Though several approaches were proposed to automate the tasks in the university campus and also image/video analysis based techniques [3, 4, 6, 20, 49] as mentioned in Section 1.1 for interpreting the student engagement, emotion, etc., but less focus on recognition and localization of students' actions inside the computer enabled laboratories in the university campus.

Over the past few years, several models were proposed for image analysis and video analysis. In the current scenario, as it is essential to provide public safety, the organizations

can use different tools to achieve this. In many organizations and public places like a bus station, airport, shopping malls and so on, the CCTV cameras are used to keep track of human activity. In the current scenario, these videos are manually analyzed to detect any type of events, behavior, and actions. Hence, it requires human resources, and also it is a time-consuming process. So several researchers were working on interpreting the contents in videos from CCTV cameras. The goal of video analysis can be of different types such as content-based video retrieval [9], activity summarizing [12], scene/event detection [38], and human action/behavior/activity recognition etc.

Many approaches were developed for video-based HAR. Many researchers proposed action recognition by considering a sequence of frames [14]. HAR can be classified based on the features used for classification; depth feature-based approaches were proposed in [11, 28]. In [32], authors proposed spatiotemporal feature-based action recognition; multi-feature based action recognition is proposed in [26]. Authors in [47], proposed a deep neural network based approach for human pose estimation by extracting the human joint positions from the depth images. A genetic algorithm combined with annealed particle filter based approach is suggested by [46] for human body pose estimation from the surveillance videos.

As a video is a consecutive set of images, hence video analysis can be considered as an extension of the image analysis where we consider the classification of images, object detection, and localization and image segmentation, etc. In image analysis, features from the still image are processed to extract the image content. A number of deep neural network [23] models like Fast Region-based Convolutional Neural Network (FRCNN)[22], YOLO[41], Single Shot Detector (SSD) [33] have been proposed for object detection and tracking.

There are many human actions that can be detected using a single image instead of using a sequence of frames that require more storage, processing power, and time-consuming. So some researchers proposed HAR using still images such as [21, 31, 51, 52], and [20]. In still image-based action recognition, researchers use the features available in a single image for action recognition. The summary of the study on still image-based action recognition is given in Table 1.

As mentioned in Table 1, several approaches by using different types of features are available for action recognition based on still images. Most of the methods available in literature recognize single action in an image. In the proposed work, we need to recognize and localize different actions performed by the students in the single image itself. Recently, many methods were proposed for object detection and tracking. It is clear from the literature survey that YOLOv3 is a faster and accurate object detector, and works well for real-time predictions. Training a new deep learning model from scratch requires a large amount of data, high computational resources, and the long duration of training. Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task[43]. Several neural network models, pretrained using a large image dataset can be used for a different kind of related application using transfer learning. HAR using transfer learning is introduced in [43], which uses a Convolutional Neural Network (CNN), Support Vector Machine (SVM), K-nearest neighbor technique to recognize human actions. The method described in [39] gives an idea of transfer learning on YOLOv3.

From the literature survey, it is clear that there is a need for a system to recognize and localize student actions inside the computer laboratories of the university campus. Further, we need a domain-specific action dataset to recognize human actions in different domains. If we consider smart classrooms, some research is done on automating some of the tasks in classrooms such as attendance management by face recognition, classroom scheduling,

Table 1 Summary of related work

Author	Method	Merits	Demerits	Dataset Used	Actions Used Used	Performance Metrics
Zheng et al. [52]	Combination of two classifiers, namely: poselet based action classifiers, learned using Poselet features, and context-based classifiers learned on contextual details are used for recognizing the human actions in still images. The method considers the probability outputs of two classifiers for each action.	Makes the use of combination of two classifiers. Good performance while recognizing some categories of the actions.	Actions used are not suitable for proposed work, so they need to be trained using actions required by the proposed method.	PASCAL VOC 2010, PASCAL VOC 2011, Willow Action Dataset	Phoning, Playing Instrument, Reading (more than 10 Actions)	mean Average Precision (mAP).
Zhang et al. [51]	Considers human poses and interaction with objects in the scene. Initially, object proposals are generated using selective search, that is decomposed into finer-grained object parts and used in finding human-object interaction regions. Features are encoded from these regions for action label prediction.	Addresses challenges in annotation efforts.	Needs improvement in the extraction of human-object interactions of different actions.	VOC 2012, STANFORD40, WILLOW 7-ACTION	Jumping, Phoning, Reading, Riding Bike and so on(More than 10 actions)	mean Average Precision (mAP)
Eweiwi et al. [20]	Action recognition in still images based on learning action specific region of interest. Optical flow fields showing human activities are extracted from video sequences to learn about salient regions for action recognition.	Demonstrates video-based feature selection for classification towards action recognition in still images.	The best recognition performance (55.2%) is achieved. Actions are not suitable for the proposed work.	Weizmann and KTH (To Learn), H3D and VOC2011 (Evaluate)	Bend, Clap, Jack, Punch, Run, Walk, Wave	Accuracy
Li et al. [31]	Uses Human-Object Interaction (HOI) for action recognition by image hierarchical representation for still action recognition. It models HOI layouts by hierarchical image representation.	Have the potential to recognize general actions and does not rely on annotations like human joint locations.	In some of the fine-grained actions, recognition performance degrades. Most of the actions are not related to the proposed work.	Sports, PPMI, extended PPMI, Standford40	Cricket batting, Cricket bowling, tennis forehand, tennis serve, People interacting with bassoon, flute and so on (more than 15 Actions)	Accuracy

augmented reality, and so on [1, 2, 25]. In many of the universities, the laboratories are made available round the clock to improve the research activity and learning in the campus. As it is difficult to have the staff, to monitor students in the laboratories all the 24 hours, it is essential to have an automated system to track students' actions in the laboratories. From the literature survey, it is also clear that no dataset is available for the actions performed by the students in the computer laboratories. So the main objective of the proposed work is to build a model for localization and recognition of students' actions inside the computer laboratory of the university campus and thus creating the STUDENT ACTION dataset.

3 Dataset creation

Our proposed model is intended to recognize students' actions inside the computer laboratory using a single image frame. The model should be able to identify actions by considering the gesture, posture, head position of the person, and the items in the surrounding. Further, we develop the STUDENT ACTION dataset with 688 image frames collected from various CCTV cameras deployed in computer laboratories in the smart campus. Some of the images used in the dataset are shown in Fig. 1. Usually, the performance of deep neural networks

(a) Image-1 (b) Image-2

(c) Image-3

Fig. 1 Sample images from the created dataset

can be significantly improved by training it with a huge amount of data. It is evident from the results of [50], and [27], data augmentation benefits in classification and object detection tasks. To improve the recognition performance of the proposed methodology, the model is trained with images of different quality. Using different augmentation techniques, dataset size is increased to 6500 images. Figure 2 shows the sample of augmented images.

For the proposed system, the created STUDENT ACTION dataset contains the class labels related to the standards mentioned in the departmental rules and regulations committee of Indian engineering laboratories. The dataset contains different action labels, as listed in Table 2.

Each image frame from the CCTV cameras deployed in computer laboratories contain different types of human actions in different parts of the image. Most of the students in

(a) Image-1a (b) Image-1b (c) Image-1c

(d) Image-2a (e) Image-2b (f) Image-2c

(g) Image-3a (h) Image-3b (i) Image-3c

Fig. 2 Sample augmented images frames

Table 2 Actions used in the dataset

Action Label	Definition
Discussion	Two or more persons sitting or standing together and facing to each other.
Engaged	A person looking at book or person looking at monitor or looking at keyboard and hands in keyboard.
Sleeping	A person bent towards the computer table and kept head on the table.
Eating	A person holding something in hand close to mouth or holding bottle in hand or near the mouth.
Using_Smart_Phone	A person looking at smartphone in hand.

laboratories will be working on computers or reading books, so we get more samples for the action label *Engaged*. The total number of samples for different action labels among the images is shown in Table 3.

The newly created STUDENT ACTION dataset contain images as well as an annotation text file. The annotation text file contains information about each image content, such as the action label and the coordinates of the location in the image where that action appears. The details of the annotation are given in Section 4.1.4. The dataset is annotated by three different annotators and follows the Gold Standard Study [18, 36] where participants (research scholars), novice judges (annotators who are not from the domain of learning technologies) and expert judges (faculty members) are the three Gold standards used for annotation. Several annotators were used in the study, and at least three different annotators annotated each image frame. The annotator's degree of agreement is measured using Kappa coefficient, and the details are given in Section 5.2.3.

4 The proposed architecture

The detailed architecture of the proposed methodology is depicted in Fig. 3. It mainly consists of the following subsystems:

– Dataset Preparation
– Training, and
– Action Recognition

In the dataset preparation subsystem, a new dataset is created for the proposed work. In the next subsystem, the pretrained YOLOv3 model is trained using the newly created dataset for

Table 3 Distribution of actions among the frames

Action Label	No. of Samples
Discussion	5034
Engaged	40052
Sleeping	6916
Eating	2092
Using Smart Phone	768

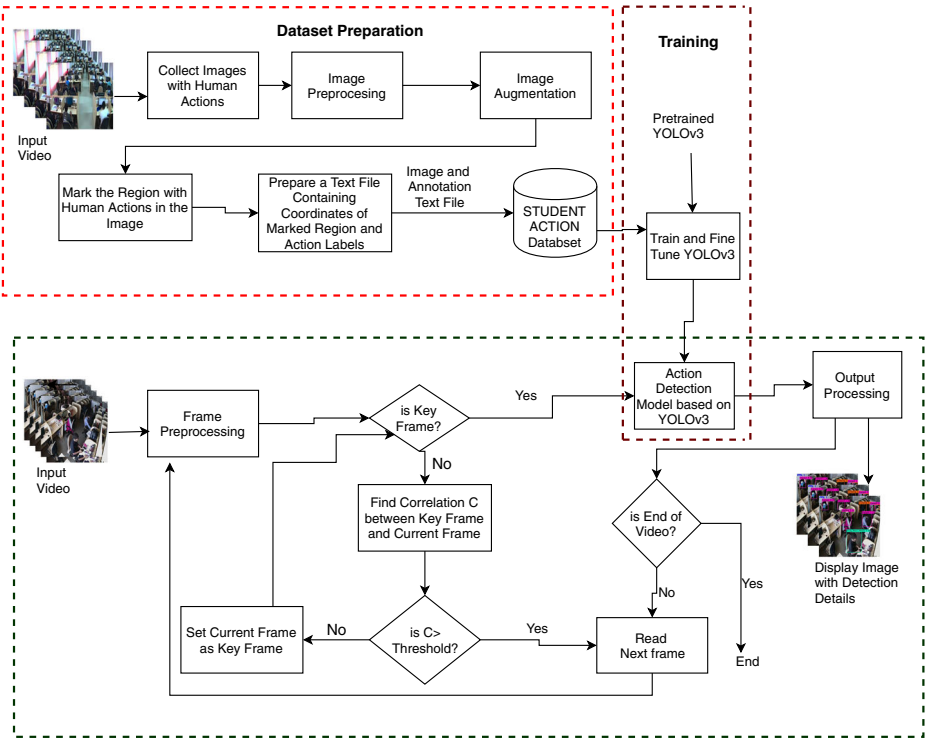


Fig. 3 Proposed architecture

the proposed work. Finally, the newly trained model is used for student action recognition. Further, since the content of continuous image frames from a video are similar to each other, we used an approach to decrease the number of frames to be processed. A detailed explanation of the proposed methodology is given below.

4.1 Dataset preparation

As the proposed system is designed for monitoring the student actions inside the computer laboratory of smart campus, we need to define an appropriate dataset that best describes student actions in the specified domain. As our purpose is to recognize actions from the still images based on gesture, posture, the head position of human, and also items nearby the human, we concentrate on only the images with human actions. So for the proposed work, the STUDENT ACTION dataset is created using the following modules.

4.1.1 Collect images with human actions

Frames from videos captured by CCTV cameras of computer laboratories are used to prepare the dataset. Image frames are extracted from videos of CCTV cameras. Image frames that contain the human actions in it are manually selected and annotated; remaining image frames are discarded.

4.1.2 Image preprocessing

As our proposed system focus on student actions, some part of the image is cropped to remove portion along the four sides of an image that does not contribute to human action. Then all images are resized into 416X416 pixels.

4.1.3 Image augmentation

Since the system is developed to monitor the students' activities in an indoor environment, especially inside the computer laboratory, we get video frames with fixed resolution. To improve the performance of the proposed system, the number of images in the dataset is increased by the image data augmentation techniques, as mentioned in Table 4.

4.1.4 Data annotation

Data annotation is the task of labeling the data. In this, we need to detect actions, identify the type of action, and find the exact location wherein the image, the action found. So, as the proposed system's primary intention is to recognize and localize the student actions inside the computer laboratory, we need to annotate the training and testing data with appropriate action labels and bounding boxes. A single image may contain different types of actions in different locations; these are recognized manually. The action labels and bounding boxes for identified actions are generated using the technique described in [24]. The coordinates of bounding boxes for each action labels in each frame are written to the annotation text file along with the action label. The annotation text file, along with the images, is given to the next step for training. The annotation format used in the proposed method is as follows:

One row for one image;
Row format: image-file-path box1 box2 ... boxN
Box format: xmin,ymin,xmax,ymax,action_label

4.2 Training

The pretrained YOLOv3 model is configured for the proposed work. Using pretrained weights of YOLOv3 and the dataset prepared in the previous step; the YOLOv3 model is

Table 4 Type of augmentations used

No.	Augmentation Technique	Description
1	GaussianBlur	It is a type of image-blurring filter that uses a Gaussian function for calculating the transformation to apply to each pixel in the image.
2	MedianBlur	The central element of the image is replaced by the median of all the pixels in the kernel area.
3	Bilateral Filter	Intensity of each pixel is replaced with a weighted average of intensity values from nearby pixels.
4	Dilation	Morphological image transformation.
5	Changing Contrast	Image with modified contrast value.

trained and fine-tuned for the proposed action recognition task. A new set of weights generated is utilized for the action recognition step in the proposed method. During the training, k-fold cross-validation method is used to improve the learning by the model.

4.2.1 YOLOv3

From the literature survey, undoubtedly, YOLOv3 is one of the fast, efficient object detectors; and has a very good real-time performance. Originally YOLOv3 is an improvement over YOLO[41]. YOLO divides the input image into an $S \times S$ grid. One object can be associated with a grid cell. It can predict a fixed number of bounding boxes, where each bounding box is associated with one box confidence score. So, the information to be predicted for each bounding box contains five values *(x, y, w, h, box confidence score)*, where the *(x, y)* coordinates represent the center of the box relative to the grid cell location and *w, h* are box dimensions, and the box confidence score says how likely the box contains an object and how accurate is the bounding box. Finally, The class confidence score for each prediction box is computed as the product of the box confidence score and conditional class probability.

Figure 4 shows the architecture of YOLOv3 with darknet-53 feature extractor [42]. The model uses successive $3x3$ and $1x1$ convolutional layers. YOLOv3 uses different types of layers, namely convolutional layers, shortcut layers, route layers, and YOLO detection layer. Where the shortcut layer provides the skip connection, the upsampling layer upsamples the feature maps from the previous layer. Route layer has an attribute called layer with one or two values, if it has one value then outputs feature maps of layer indexed by the value, otherwise concatenated feature maps of layers indexed by its values. Finally, the YOLOv3 detection layer specifies the anchors (default bounding boxes) used during detection.

YOLO Loss Function: The loss function consists of three components:

- **Classification Loss:** If an object is detected, at each cell, it is estimated as the squared error of the class conditional probabilities for each class.
- **Localization Loss:** Measures the errors in the predicted bounding box locations and sizes.
- **Confidence Loss:** Measuring the objectness of the box.

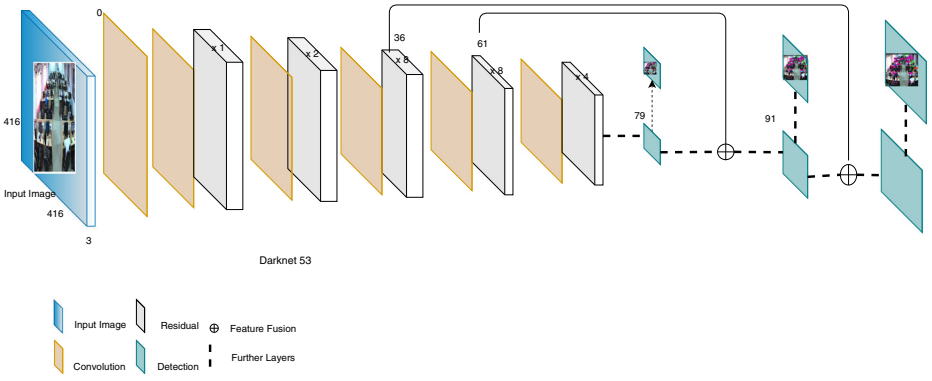


Fig. 4 YOLOv3 framework used in the proposed model

4.3 Action recognition

Algorithm 1 Algorithm for action recognition subsystem.

Input: Video Stream from CCTV camera
Output: Image frames with action labels and bounding boxes around detected student actions
Initialize: flag = True
Extract first image frame from video stream
Preprocess the image frame and set it as keyframe and current_frame
while *flag* **do**
  if *keyframe and current_frame are same* **then**
    Detect actions in current_frame using fine tuned action detection model
    Display image along with detection labels and bounding boxes around recognized actions
    if *end of video stream* **then**
      flag = False
  else
    find Template Match between keyframe and current_frame and set it as C
    if $C \leq threshold$ **then**
      keyframe = current_frame
    else
      skip current_frame
      Extract and preprocess next frame and set it as current_frame
    end
  end
end

The task of this subsystem is to analyze the video from CCTV cameras of the computer laboratories. It is responsible for the recognition and localization of the actions performed by the students.

The model obtained after training is capable of recognizing and localizing student actions in a single image frame. As a consecutive set of image frames in a video are very similar to each other, in the proposed work to speed up the video analysis, we try to reduce the number of frames to be processed using template matching. We find a similarity between two frames based on template match value. If the template match value is higher than the threshold, skip that frame and read the next frame. In the proposed system, we set the threshold for template matching as 0.97. During experiments, it is observed that time taken for template matching is less than time taken to process the frame for action detection. It is noted that in the system in which experiments are performed, it takes on an average 0.099 seconds for template matching at the same time, 0.90 seconds are required for action detection. It is also observed that actions in skipped frames are present in frames that are not skipped. So in this way, the number of frames that need to be processed for video analysis is greatly reduced. The steps used in action recognition are given in Algorithm 1. Algorithm 1 runs for $O(n)$ times where n is the number of image frames in the video.

In Algorithm 1, the template match is calculated using (1) as the parameter and described by the method given in [35].

$$R(x, y) = \frac{\sum_{x', y'}(T(x', y') \cdot I(x + x', y + y'))}{\sqrt{\sum_{x', y'} T(x', y')^2 \cdot \sum_{x', y'} I(x + x', y + y')^2}} \tag{1}$$

Where I, T and R represent the image, template, and result, respectively. The parameters (x, y) are the pixel positions. The parameter x' can vary from zero to width of the template, and y' can vary from zero to height of the template. In the proposed system, I is taken as the current keyframe, and T is considered as the current frame extracted from the video.

5 Experimental results and analysis

The details of observations made during the training and testing of the proposed model are described in this section.

5.1 Training

The proposed model is trained using the k-fold cross-validation technique. The entire dataset is divided into ten folds where each fold contains 650 images. Each of the ten folds is tested while the remaining nine folds are used for training. The model is trained for 35 epochs using two Tesla M40 GPUs with 256 GB RAM and 256 GB unallocated space. Figure 5 shows the training and validation loss comparisons with all ten folds in a maximum of 35 epochs, and it is observed that the loss is within the range of 30 to 35 showing that there are no significant issues like overfitting or any irregular results due to skewed dataset in the proposed method.

5.2 Testing

The proposed system for student action recognition using still images provides promising results. The proposed model is tested on each of the ten folds. Figures 6, 7, and 8 show some of the sample input and the corresponding output from the frames used for testing. There are three sample input images shown here. In all these images, students are involved in different

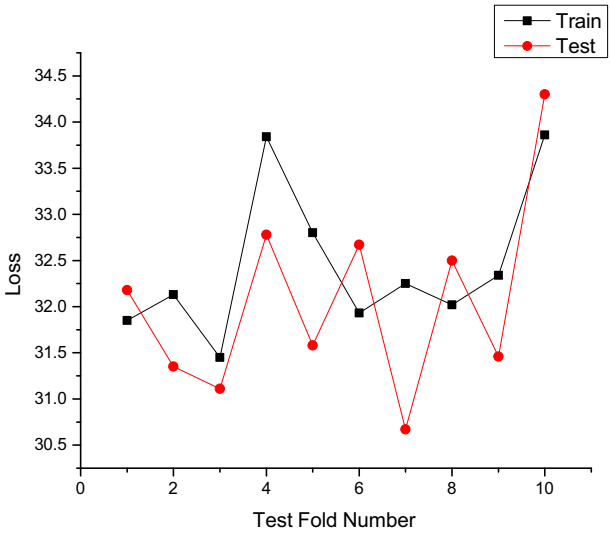
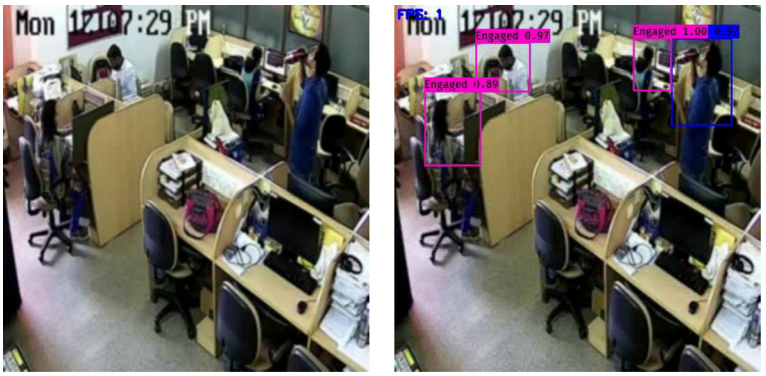


Fig. 5 Training and validation loss



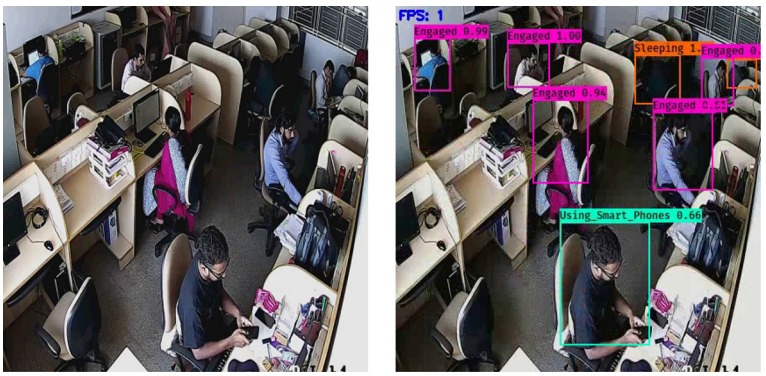
(a) Input _image-1 (b) Output _image-1

Fig. 6 Sample input and output -1

actions inside the computer laboratory. In *Input_image-1*, three students are *Engaged*; that is, the students are working on the computer, and one student is holding the bottle near the mouth, so the action is *Eating*. *In Input_image-2*, many students are *Engaged*, one student is *Using_smart_phone*, and two students are *Sleeping*. Similarly, *Input_image-3* shows samples for *Discussion*, *Sleeping*, and *Engaged* actions. The corresponding output images from the proposed model show the detected action, display the action label along with bounding boxes, and the score for action detection.

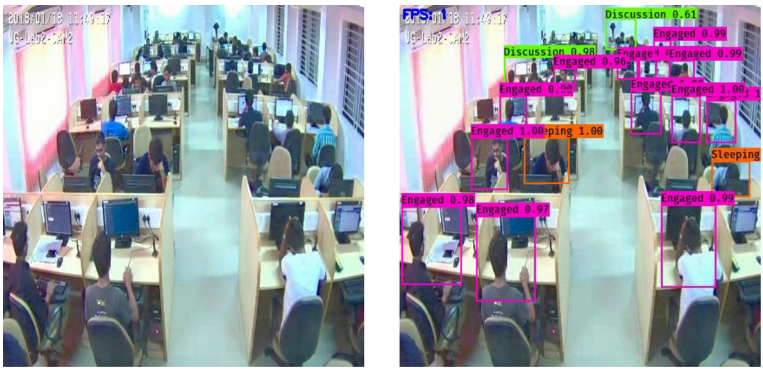
5.2.1 Ground-truth information of images used for testing

In the following part of the paper, the details of the observed results are discussed. Figure 9 shows the ground-truth information of the set of frames in one of the folds as a sample for better understanding of the class label distribution. This sample has 650 frames with a total of 3984 *Engaged*, 511 *Discussion*, 706 *Sleeping*, 218 *Eating*, and 64 *Using_Smart_Phone*.



(a) Input _image-2 (b) Output _image-2

Fig. 7 Sample input and output -2



(a) Input _image-3 (b) Output _image-3

Fig. 8 Sample input and output -3

5.2.2 IoU (Intersection Over Union):

Since the proposed system needs to detect and locate all the actions in a single image frame. We also considered IoU (Intersection Over Union) while evaluating the system. IoU is an evaluation metric used while measuring the accuracy of an object detector on a particular dataset. To measure IoU, we used the ground truth bounding boxes and predicted bounding boxes. IoU is mathematically defined as follows:

$$IoU = \frac{Area\ of\ Overlap\ of\ Two\ Bounding\ Boxes}{Area\ of\ Union\ of\ Two\ Bounding\ Boxes} \tag{2}$$

5.2.3 Kappa coefficient

Originally it was introduced by [15] to measure the inter-rator agreement on nominal scales. In the proposed methodology, we used the Kappa coefficient to measure the agreement between the annotators in recognizing and locating the actions in image frames used in the dataset. The value of Kappa value varies from 0-1. Here the value 1 indicates perfect

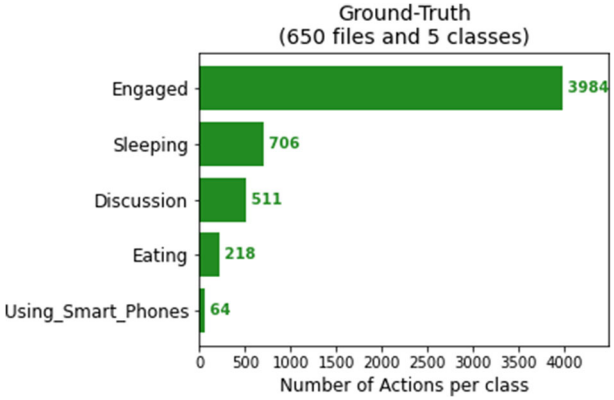


Fig. 9 A sample distribution of student actions in a fold of cross-validation

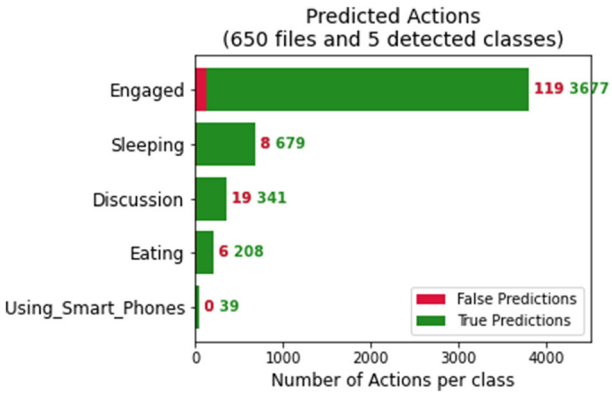


Fig. 10 Information about detected actions

agreement between the annotators, and 0 indicates no agreement.We adapted the formula for kappa coefficient from [45], as shown below:

$$K = \frac{N \sum_{i=1}^{n} m_{i,i} - \sum_{i=1}^{n} G_i * C_i}{N^2 - \sum_{i=1}^{n} G_i * C_i} \tag{3}$$

Where: G_i and C_i indicate true values and predicted values belonging to class i. Variable m indicates confusion matrix.

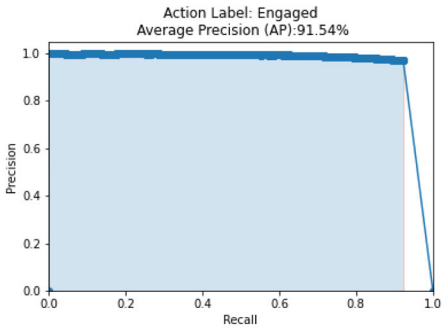
The detection result from the proposed system using IoU = 0.45 is shown in Fig. 10. Table 5 provides the details about True Positive (TP) and False Positive (FP) obtained for each action label during testing.

To compare annotations by different annotators, in the proposed methodology, C and G are considered as annotations by annotator1 and annotator2, respectively. For annotator 3, C contains the average of annotator 1 and 2, and G contains annotator3.
i is the total number of action labels.
N is the total maximum of the total number of annotations by both annotators. $m_{i,i}$ is the number of values both annotators annotated as action label i.
C_i is the number of values belonging to action label i according to annotator1.
G_i is the number of values belonging to action label i according to annotator2.

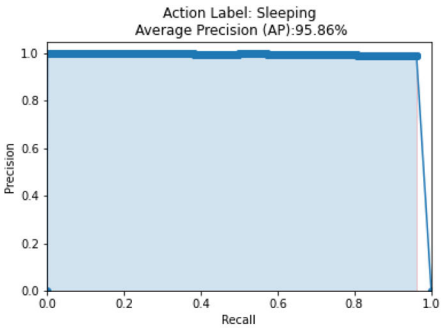
We observed that the Kappa coefficient of annotations is 0.65. The interpretation of different Kappa values can be seen in [30]. Here the range of 0.60 to 0.80 indicates substantial agreement between the two observers. As we got Kappa value greater than 0.6, we concluded that there is a good agreement between the annotation by the annotators though there is no perfect agreement.

Table 5 TP and FP found for action labels while testing

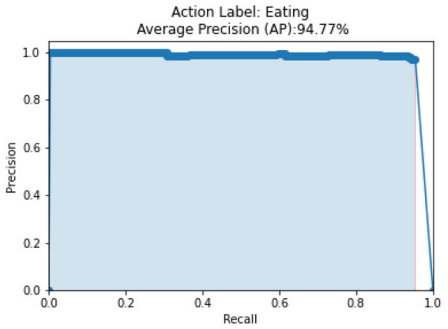
Action Label	Ground_Truth	Total Detected	True Positive (TP)	False Positive (FP)
Discussion	511	360	341	19
Engaged	3984	3796	3677	119
Sleeping	706	687	679	8
Eating	218	214	208	6
Using_ Smart_Phone	64	39	39	0



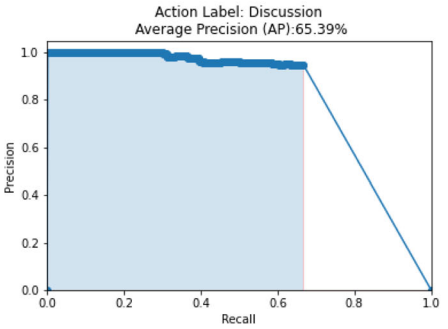
(a) Action: Engaged



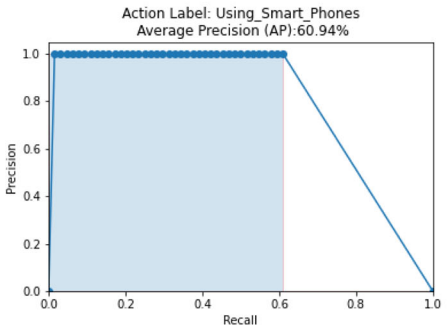
(b) Action: Sleeping



(c) Action: Eating



(d) Action: Discussion



(e) Action: Using _smart _phones

Fig. 11 Precision/recall curves obtained for different actions

5.2.4 Analysis of detected actions

5.2.5 Precision/recall curves

Precision and recall are the two important measures used to evaluate the classifiers. Precision measures the fraction of relevant instances among the retrieved instances. Recall measures the fraction of relevant instances retrieved over the total amount of relevant

instances. Equations (4) and (5) show the method used to calculate precision and recall, respectively.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{4}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{5}$$

Precision-Recall (PR) curve gives a more informative picture of an algorithm's performance when dealing with highly skewed datasets [17]. It summarizes the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds, and it is better to use this in the case of moderate to large class imbalance[5]. We obtain PR curves for all the five different actions by mapping each detection to a ground-truth class instance as in [7] and [8]. Figure 11 shows the different PR curves for all the five different actions obtained during testing. The average precision is 91.54%, 95.86%, 94.77%, 65.39%, and 60.94% for *Engaged*, *Sleeping*, *Eating*, *Discussion*, and *Using_Smart_Phone* actions, respectively. The performance of the proposed model is evaluated using Average Precision (AP) for all action classes and by mean Average Precision (mAP). AP for all actions and mAP obtained during testing is as shown in Fig. 12. The mAP obtained during testing is 81.70%.

The experimental result obtained by the proposed system shows that it performs well in recognition and localization of student actions with which the model is trained with more number of samples than actions with less number of samples used for training.

Finally, we used template matching in action recognition subsystem to process the video with lesser processing time. The steps are given in Algorithm 1. After several tests, the threshold for template matching is empirically set to 0.97. If the current frame is similar to the keyframe, then the current frame is skipped from action detection; instead, keyframes' actions are considered. The details of the test on three different CCTV camera videos are shown in Table 6. From the result, it is clear that the template matching operation is less time consuming than action detection. Also, the percentage of actions added or deleted from the skipped frames when compared to the keyframes are very less. Also, the effect of removed or added actions is negligible while considering the entire video. Thus, it significantly reduces the time required to process the complete video. So, we can also significantly reduce the size of the video for future reference by eliminating unnecessary frames. Table 6

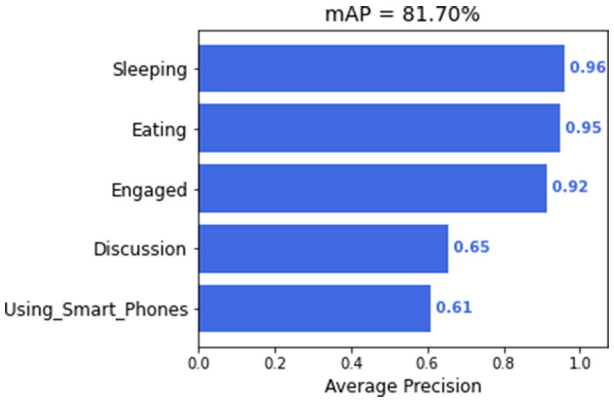


Fig. 12 mAP obtained during testing

Table 6 Comparison of video processing with and without template matching

Name	T1.mp4	T2.mp4	T3.mp4
Total number of frames	500	650	1001
Number of skipped frames with template matching	429	298	982
Number of frames considered for action detection with template matching	71	352	19
Average time for action detection per frame (seconds)	0.9	0.9	0.9
Average time for template matching per frame (seconds)	0.03	0.03	0.08
Total time taken to process video with template matching (seconds)	84.30	352.59	108.18
Total time taken to process video without template matching (seconds)	464.35	610.86	955.41
Extra actions added w.r.t first keyframe (%)	0	1	0
Extra actions added w.r.t second keyframe (%)	3	0	0
Actions deleted w.r.t first keyframe (%)	3	0	2.04
Actions deleted w.r.t second keyframe (%)	1.5	0	0

gives details about the percentage of other actions added and actions missed with respect to the first two keyframes obtained and the time required to process the complete video with and without template matching on the computer system which was used for testing.

5.3 Limitation

Though the proposed system gives promising results in detecting and locating the student actions, the proposed model did not consider students from other laboratories which are not computer-enabled. The proposed methodology does consider mixed human actions like engaged and eating, eating and discussing, eating and using_smart_phone, and so on. The proposed model did not consider posed expressions in the dataset. Hence there will be an imbalance in the dataset, which affects the performance of the proposed model.

6 Conclusion and future work

In this, a real-time student monitoring system using transfer learning on YOLOv3 is proposed. The proposed model gives a promising result in the recognition and localization of the students' actions inside the computer laboratory in the university campus. The model also shows that a certain set of human actions are recognized using a single image frame. The proposed model also achieved a reduction in the time required for video analysis by reducing the number of frames to be processed. So the proposed system can be used for real-time monitoring of student activities inside the computer laboratory.

As a further improvement to the system, it can be trained with more types of actions and also improve the classification and localization accuracy. We can also further improve the model to identify the person involved in a particular action.

Ethics Statement Authors have obtained all ethical approvals from the Institutional Ethics Committee (IEC) of National Institute of Technology Karnataka Surathkal, Mangalore, India and a written consent was also obtained from the human subjects.

References

1. Ashwin T, Guddeti RMR (2020) Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. Educ Inf Technol 25(2):1387–1415
2. Ashwin TS, Guddeti RMR (2019) Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues. IEEE Access 7:150,693–150,709
3. Bian C, Zhang Y, Yang F, Bi W, Lu W (2019) Spontaneous facial expression database for academic emotion inference in online learning. IET Comput Vis 13(3):329–337
4. Bosch N, D'Mello S (2019) Automatic detection of mind wandering from video in the lab and in the classroom. IEEE Trans Affect Comput 1–1. https://doi.org/10.1109/TAFFC.2019.2908837
5. Brownlee J How and when to use roc curves and precision-recall curves for classification in python,. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/. Accessed: 14 05 2019
6. Candra Kirana K, Wibawanto S, Wahyu Herwanto H (2018) Facial emotion recognition based on viola-jones algorithm in the learning environment. In: 2018 International seminar on application for technology of information and communication, pp 406–410
7. Cartucho: map (mean average precision), https://github.com/Cartucho/mAP. Accessed: 12-06-2020
8. Cartucho J, Ventura R, Veloso M (2018) Robust object recognition through symbiotic deep learning in mobile robots. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 2336–2341
9. Castañón G, Elgharib M, Saligrama V, Jodoin P (2016) Retrieval in long-surveillance videos using user-described motion and object attributes. IEEE Trans Circuits Sys Video Technol 26(12):2313–2327
10. Chamba L, Aguilar J (2016) Design of an augmented reality component from the theory of agents for smart classrooms. IEEE Lat Am Trans 14(8):3826–3837
11. Chaudhary S, Murala S (2019) Depth-based end-to-end deep network for human action recognition. IET Comput Vis 13(1):15–22
12. Cheng H, Liu Z, Zhao Y, Ye G, Sun X (2014) Real world activity summary for senior home monitoring. Multimed Tools Appl 70(1):177–197. https://doi.org/10.1007/s11042-012-1162-5
13. Chintalapati S, Raghunadh MV (2013) Automated attendance management system based on face recognition algorithms. In: 2013 IEEE International conference on computational intelligence and computing research, pp 1–5
14. Chou K, Prasad M, Wu D, Sharma N, Li D, Lin Y, Blumenstein M, Lin W, Lin C (2018) Robust feature-based automated multi-view human action recognition system. IEEE Access 6:15,283–15,296
15. Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46. https://doi.org/10.1177/001316446002000104
16. Conte D, Foggia P, Percannella G, Tufano F, Vento M (2010) A method for counting moving people in video surveillance videos. EURASIP Journal on Advances in Signal Processing 2010(1):231–240
17. Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, pp 233–240
18. D'Mello S, Picard RW, Graesser A (2007) Toward an affect-sensitive autotutor. IEEE Intell Syst 22(4):53–61
19. Du S, Meng F, Gao B (2016) Research on the application system of smart campus in the context of smart city. In: 2016 8th International Conference on Information Technology in Medicine and Education (ITME), pp 714–718
20. Eweiwi A, Cheema MS, Bauckhage C (2015) Action recognition in still images by learning spatial interest regions from videos. Pattern Recogn Lett 51(C):8–15
21. Ghazal S, Khan US (2018) Human posture classification using skeleton information. In: 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp 1–4
22. Girshick RB (2015) Fast r-cnn. arXiv:1504.08083
23. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge. http://www.deeplearningbook.org
24. Gu J (2019) Bbox-label-tool. https://github.com/jxgu1016/BBox-Label-Tool-Multi-Class. Accessed 02 Aug 2019

25. Gupta SK, Ashwin T, Reddy Guddeti RM (2018) Cvucams: Computer vision based unobtrusive classroom attendance management system. In: 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), pp 101–102
26. Huang M, Su SZ, Zhang HB, Cai GR, Gong D, Cao D, Li SZ (2018) Multifeature selection for 3d human action recognition. ACM Trans Multimedia Comput Commun Appl 14(2):45:1–45:18
27. Jo H, Na Y, Song J (2017) Data augmentation using synthesized images for object detection. In: 2017 17th International Conference on Control, Automation and Systems (ICCAS), pp 1035–1038
28. Kamel A, Sheng B, Yang P, Li P, Shen R, Feng DD (2018) Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 1–14
29. Kim Y, Soyata T, Behnagh RF (2018) Towards emotionally aware ai smart classroom: Current issues and directions for engineering and education. IEEE Access 6:5308–5331
30. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174. http://www.jstor.org/stable/2529310
31. Li R, Liu Z, Tan J (2018) Reassessing hierarchical representation for action recognition in still images. IEEE Access 6:61,386–61,400
32. Li W, Nie W, Su Y (2018) Human action recognition based on selected spatio-temporal features via bidirectional lstm. IEEE Access 6:44,211–44,220
33. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision – ECCV 2016. Springer International Publishing, Cham, pp 21–37
34. Monkaresi H, Bosch N, Calvo RA, D'Mello SK (2017) Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans Affect Comput 8(1):15–28. https://doi.org/10.1109/TAFFC.2016.2515084
35. OpenCV -Object Detection: Opencv -object detection, https://docs.opencv.org/3.4.3/df/dfb/group__imgproc__object.html. Accessed: 12-04-2019
36. Picard RW (2000) Affective computing. MIT Press, Cambridge
37. Popoola OP, Wang K (2012) Video-based abnormal human behavior recognition—a review. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(6):865–878
38. Poulisse GJ, Patsis Y, Moens MF (2014) Unsupervised scene detection and commentator building using multi-modal chains. Multimedia Tools and Applications 70(1):159–175. https://doi.org/10.1007/s11042-012-1086-0
39. qqwweee: keras-yolo3. https://github.com/qqwweee/keras-yolo3. Accessed: 05-01-2019
40. Ramezani M, Yaghmaee F (2016) A review on human action analysis in videos for retrieval applications. Artif Intell Rev 46(4):485–514
41. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 779–788
42. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv:1804.02767
43. Sargano AB, Wang X, Angelov P, Habib Z (2017) Human action recognition using transfer learning with deep representations. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp 463–469
44. Sari MW, Ciptadi PW, Hardyanto RH (2017) Study of smart campus development using internet of things technology. IAES International Conference on Electrical Engineering, Computer Science and Informatics IOP Conf Series: Materials Science and Engineering 190(2017):012032. https://doi.org/10.1088/1757-899X/190/1/012032
45. Sivabalan K, Ramaraj E (2020) Shortwave infrared-based phenology index method for satellite image land cover classification. In: Das K, Bansal J, Deep K, Nagar A, Pathipooranam P, Naidu R (eds) Soft computing for problem solving advances in intelligent systems and computing. Springer 1057. https://doi.org/10.1007/978-981-15-0184-5_75
46. Szczuko P (2014) Genetic programming extension to apf-based monocular human body pose estimation. Multimedia Tools and Applications 68. https://doi.org/10.1007/s11042-012-1147-4
47. Szczuko P (2019) Deep neural networks for human pose estimation from a very low resolution depth image. Multimedia Tools and Applications 1–21. https://doi.org/10.1007/s11042-019-7433-7
48. Wang C, Li X, Wang A, Zhou X (2017) A classroom scheduling service for smart classes. IEEE Trans Serv Comput 10(2):155–164
49. Whitehill J, Serpell Z, Lin Y, Foster A, Movellan JR (2014) The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Trans Affect Comput 5(1):86–98. https://doi.org/10.1109/TAFFC.2014.2316163
50. Wong SC, Gatt A, Stamatescu V, McDonnell MD (2016) Understanding data augmentation for classification: when to warp? arXiv:1609.08764

51. Zhang Y, Cheng L, Wu J, Cai J, Do MN, Lu J (2016) Action recognition in still images with minimum annotation efforts. IEEE Trans Image Process 25(11):5479–5490
52. Zheng Y, Zhang Y, Li X, Liu B (2012) Action recognition in still images using a combination of human pose and context information. In: 2012 19th IEEE International Conference on Image Processing, pp 785–788

M. Rashmi received the M.Tech degree in Computer Science and Engineering from Visveswaraya Technological University, Belgaum, India, in 2014. She is currently pursuing the Ph.D. degree with the Department of Information Technology, National Institute of Technology Karnataka, Mangalore, India. Her research interests include computer vision, human action/activity recognition, and smart city/campus applications. She is a student member of IEEE.

T. S. Ashwin received the B.E. degree from Visveswaraya Technological University, Belgaum, India, in 2011, and the M.Tech. degree from Manipal University, Manipal, India, in 2013. He received his Ph.D. degree from National Institute of Technology Karnataka Surathkal, Mangalore, India. He is currently working as an associate professor at SJEC Vamanjoor Mangalore. He has more than 35 reputed and peer-reviewed international conferences and journal publications. His research interests include multi-modal affective content analysis, emotional, behavior and cognitive student engagement analysis, recommender systems, auto tutors, game-based learning, smart classroom environments, and computer vision applications. He is a Member of the IEEE and ACM.

Ram Mohana Reddy Guddeti received the B.Tech. degree from Sri Venkateswara University, Tirupati, India, in 1987, the M.Tech. degree from the IIT Kharagpur, Kharagpur, India, in 1993, and the Ph.D. degree from The University of Edinburgh, U.K., in 2005. He is currently the Senior Professor in the Department of Information Technology, National Institute of Technology Karnataka, Mangalore, India. He has more than 200 research publications in reputed and peerreviewed international journals, conference proceedings, and book chapters. His research interests include affective computing, big data and cognitive analytics, bio-inspired cloud and green computing, the Internet of Things and smart sensor networks, social multimedia, and social network analysis. He is a Senior Member of the IEEE and ACM, a Life Fellow of IETE (India), a Life Member of the Computer Society of India, and a Life Member of ISTE (India).