Smart Blended Learning Framework based on Artificial Intelligence using MobileNet Single Shot Detector and Centroid Tracking Algorithm

Abdul Wahid[1], Muhammad Fajar B[2], Jumadi M. Parenreng[3], Seny Luhriyani[4], Puput Dani Prasetyo Adi[5]
Computer Engineering Study Program, Universitas Negeri Makassar, Makassar, Indonesia[1, 2, 3]
English Department, Universitas Negeri Makassar, Makassar, Indonesia[4]
National Research and Innovation Agency (BRIN), Bandung, Indonesia[5]

***Abstract*—The Covid-19 pandemic has affected all aspects of human life and has even forced humans to shift their life habits, including in the world of education. The learning model must shift from the traditional face-to-face pattern to a modern face-to-face pattern or an asynchronous pattern with information technology-based applications. Blended learning is one of the appropriate solutions to adjust the limited face-to-face learning conditions. Blended learning can be done, for example, by scheduling learning by dividing the number of participants by 50% and entering on a scheduled basis. However, the problem is that the time and effort used are less efficient. Blended learning can also be done by conducting learning simultaneously with 50% of students in class and the remaining 50% through conferences. This concept will streamline the time and effort used. However, the problem is that there is a gap in the learning experience between students in class and students who do learning via conference. This innovative blended learning system framework is proposed to overcome these problems. The system built seeks to present an online learning experience atmosphere so that it is expected to be able to resemble an offline learning atmosphere. We created a system using camera technology and object detection that will track the movement of the teacher so that the teacher can move freely in the room without having to be stuck in front of the computer holding the conference. The algorithms used are MobileNet Single Shot Detector and Centroid Tracking. This research produces an accurate model for detecting teacher movement at a distance of 2, 4, and 6 meters with a camera installation height of 1.5 and 3 meters.**

Keywords*—Smart blended learning; mobilenet; single shot detector; convolutional neural network; centroid tracking

I. Introduction

Blended learning has been applied in higher education for several years, but there is still limited research on what affects student satisfaction in blended learning environments in higher education [1]. The implementation of blended learning at a university is evaluated by measuring student satisfaction in face-to-face sessions, independent study sessions using online learning, and the overall learning experience in a Blended Learning environment.

There are two main areas related to the blended learning environment. The first is a blend of traditional classroom learning and e-learning, and the second is synchronous and asynchronous e-learning technology [2]. This first area is the best-known form of combination seen in the amalgamation of theory and practice of instructor and student-centered learning. The second type of blended learning is the blend of technologies that give students access to synchronous and asynchronous communication and information. This is especially useful when considering the number of external students outside of campus studying at the tertiary level and the associated geographic and access issues and creating an environment accommodating cross-cultural learners.

Along with the development of artificial intelligence technology, various technologies are born to assist humans in completing a task or activity. The technology in question is capable of performing actions like a human. An example is computer intelligence replacing the human sense of sight, usually known as computer vision [3]. By utilizing camera functions supported by object detection algorithms, today's computers can intelligently carry out surveillance automatically like a human's vision.

On the one hand, the current Covid-19 pandemic requires physical interaction restrictions to prevent virus transmission [4]. These restrictions have brought significant changes in various fields, especially in education, namely the limited offline learning activities that are shifted to semi-online and even full-online learning to avoid interactions that can spread virus transmission.

One of the benefits of blended learning is that learning becomes more flexible because its implementation is not limited by distance and time. Online learning from home and offline learning in the classroom can be carried out simultaneously [5]–[9]. However, there are shortcomings, namely the limited space for teachers and students in learning activities. There is a gap between the learning experience in offline classes and online learning conditions. Therefore, a system can be developed to overcome this gap by utilizing computer vision technology, especially object detection technology. The system in question is a system that can use a camera to detect objects on the teacher so that the teacher has free space to explain like in an offline class, and the teacher's position will still focus on video because of the combination of objects detection algorithms and tracking algorithms. The benefit for students is to present a learning experience that tries to approach conditions like in an offline class.

II. Literature Study

Object tracking is essential in computer vision and widely used in human-computer interaction, surveillance, and medical imaging. In its simplest form, tracking can be defined as estimating the trajectory of an object in the image plane as it moves around the frame [10]. Object tracking has attracted significant attention because it can perform a wide range of processes, including intelligence in video surveillance, machine and human interfaces, and the field of robotics [11]. However, designing an excellent visual tracking method is still an open issue. Challenges in visual tracking problems include varying shapes and appearances of objects, occlusion, lighting changes, irregular scenes, etc.

The object tracking process consists of two stages in analyzing the video, namely detecting objects and tracking the movement of objects from frame to frame [12]. One of the tracking algorithms that can be used is the centroid tracking algorithm. The centroid tracking algorithm works by taking into account the center point (centroid) of an object in tracking [13]. Therefore, the tracking process in the centroid tracking algorithm is very dependent on the accuracy of the object detection or identification algorithm. The tracking process will be challenging if the object detection algorithm has minimal accuracy and takes a long time. An object tracking task requires an object detection process that can work in real-time and has good accuracy to avoid decision errors.

Research on object identification using a neural network has been carried out by Girshick by proposing the R-CNN architecture [14]. The results obtained are 66% accuracy with 47 seconds per image detection time. The long detection time is due to the algorithm classifying as many as two thousand proposal regions for each image. Furthermore, Girshick optimized the study [14] and proposed the Fast R-CNN architecture [15]. This algorithm can shorten the detection time to 25 times faster than R-CNN and produces an accuracy of 66.9%. The increase in detection speed occurs because the CNN process was initially carried out amounted to about two thousand times be reduced to only one time. The detection time per image only takes 2 seconds.

Research by Ren et al. [16] proposed an architecture called Faster R-CNN. The idea of Faster R-CNN is to replace the selective search algorithm to generate proposal regions used in research [14], [15], becoming a Region Proposal Network (RPN). The detection time per image can reach 0.2 seconds with an accuracy of 66.9%.

Li et al. [17] carried out subsequent research to test the performance of an architecture called MobileNet SSD. The test results show that the accuracy can reach 95% with a detection time of 0.12 seconds. MobileNet SSD combines MobileNet architecture and Single Shot Multibox Detector (SSD) that uses depthwise and pointwise convolution, reducing computation significantly [18]. MobileNet SSD can provide faster object detection performance compared to Faster R-CNN.

Rahman et al. [13] conducted research on object detection in the form of vehicles to find out which vehicles are against the current or in the opposite direction. The movement of objects is known to use the YOLO detection method and centroid tracking. Centroid tracking works well on single class objects, but if applied to multiclass objects; it allows the identity of object tracking to be exchanged. The results of this study can track vehicles against the direction of the current, which is tested on a 1280x720 video.

Distance learning is usually carried out via video conferencing. In general, video conferencing use technology assistance such as WebRTC [19]. This technology allows video, audio, or other data to be transferred to the student in real-time. In blended learning, the problem is that the video source that is usually used is placed in a static position in front of the teacher so that the teacher can only stand in front of the video source camera. In our proposed research, object detection and tracking technology are implemented with the aim that the teacher's camera can be placed far from the teacher to shoot a wide area in the classroom so that wherever the teacher moves, the video will still focus on the teacher in the center of the video.

III. Methods

Data collection is the first step in the research process—image data is produced by splitting test and training data. The test data is organized around a single object class: humans. The previously provided test data is fed into the pre-processed data. The data is pre-processed to create a more optimum feature extraction. The next step is to extract the features to obtain a value utilized as input in the following stage. MobileNet SSD is the object detection algorithm. The outcomes of the architectural trials are analyzed, and the best accuracy is chosen for use as training for future test data. As for the test data, pre-processed data is carried out, and then we proceeded with feature extraction and final testing. Workflow can be seen in Fig. 1.

A. Data Acquisition Stage

The data used as input is a real-time video recording. Aside from video recordings, the dataset was derived from open-source internet sources, specifically Open Images and YouTube videos containing things to be detected, specifically human objects.

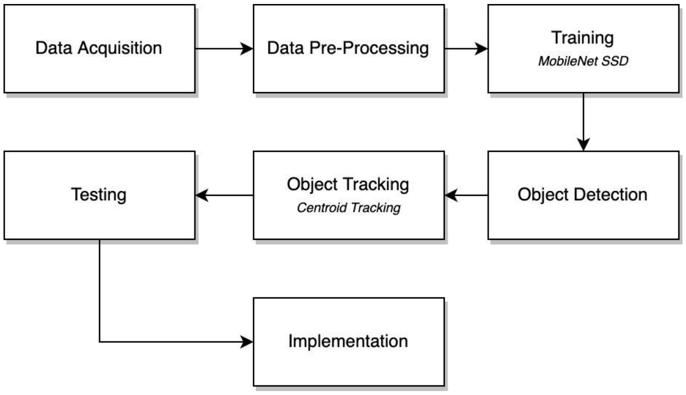

Fig. 1. Research Methodology.

B. Data Pre-processing Stage

After the input dataset was collected, the annotation process was carried out. Annotation is the process of creating labels by providing a bounding box along with the object class name in each image. In this study, the annotations were stored in a file containing information about the object class, each bounding box's coordinates, and its label. Details of the number of datasets are:

- The dataset is an image of a human class object.
- We amassed about 330 data, consisting of 227 training data and 33 test data.
- The number of annotations on the training data is 1,307, while the number of annotations on the test data is 83.

C. Training Stage

Object recognition training is carried out using an annotated image dataset. Then a convolution process is performed using the MobileNet SSD architecture to train the model weights to accurately categorize objects visible on the video camera. In the training stage, iteration was determined by a threshold indication, which takes the form of a loss function value, as seen in Fig. 2. The lower the loss value, the better the developed object detection model.

The SSD MobileNet architecture combines the MobileNet architecture as the base network and the SSD architecture as the detecting network.

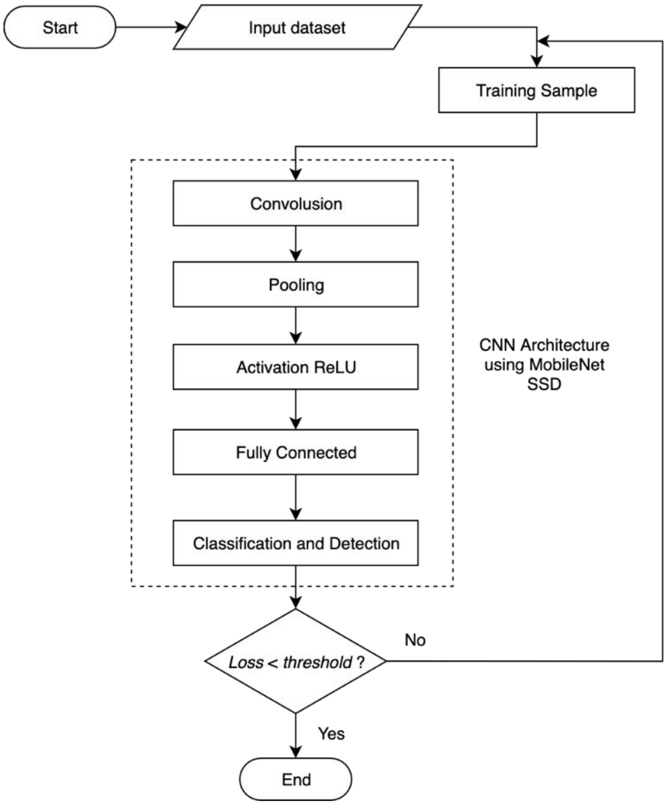

Fig. 2. Block Diagram of Training Process.

Table I depicts MobileNet's standard architecture. Column *n* specifies the number of layers, column *c* specifies the output size, and column *s* specifies the stride [20]. Before average pooling, the MobileNet architecture employs all layers, and the SSD architecture replaces the layers in the typical pooling and fully connected network.

TABLE I. MOBILENET ARCHITECTURE

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
1 × 1 × 1280	conv2d 1x1	-	k	-	

A multiscale feature map layer is employed in SSD design, which indicates that various feature map sizes are used. In general, the feature map sizes on SSDs are 512, 256, 256, and 128 [21]. Modifications were performed in this study by adding a 64 feature map, resulting in the SSD network configuration being 512, 256, 256, 128, and 64. The addition of this feature map seeks to evaluate the model's effectiveness in recognizing items of smaller size.

D. Object Detection

After the training stage is completed, the model can be used to identify the trained objects. Fig. 3 depicts the steps of object identification.

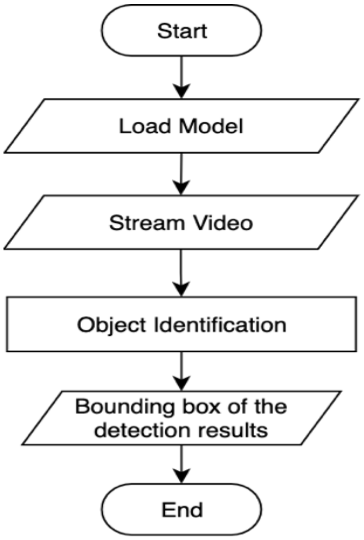

Fig. 3. Block Diagram of Object Identification Process.

E. *Object Tracking*

Object tracking is used to identify the detecting object so that its movement in a movie can be monitored from frame to frame. The object tracking algorithm used in this work is a centroid tracking approach, although it has been changed to improve performance. Tracking workflow can be seen in Fig. 4. The centroid tracking algorithm is based on the Euclidean distance between the existing centroid object and the new centroid object between frames in a video. The following are the suggested tracking steps:

- Determine the centroid at each of the bounding box locations.
- Calculate the Euclidean distance between the new and old bounding boxes.
- Update the coordinates (x, y) based on the object and class category.
- If a new detection occurs, the new object is added to the new track.
- If the tracking object has been lost, it is removed by setting a threshold for the next N frames.

The centroid tracking approach can perform effectively if the object detection model delivers correct results. The object detection model may not be accurate in the object detection process. Therefore, there is the possibility of mistakes in the form of objects that are not identified or detected but are less accurate. For example, one object is detected as two or more objects. The Non-Maximum Suppression (NMS) method is used in this work to suggest an additional way of post-process detection. NMS works to solve the problem of overlapping bounding boxes from detection results. Incorporating this process will aid the detection process in determining if the bounding box above another bounding box is still the same object or a distinct object [22]. The following are suggested steps from NMS:

- Step 2 should be repeated for each separate object class.
- Take the last index from the index list's enclosing box and add the index value to the specified index list.
- Find the greatest coordinate (x, y) for the bounding box's start and the smallest coordinate (x, y) for the bounding box's end.
- Determine the bounding box's width and height.
- Determine the overlap ratio.
- Remove from the index list any indexes with overlapping threshold values.

F. *Testing Stage*

The tracking performance testing procedure steps are carried out by applying detection and tracking algorithms on image and video testing with a variety of tests, including camera distance and height testing, frame rates at various video resolutions, and object closures. The F1 score from the confusion matrix is used to calculate performance.

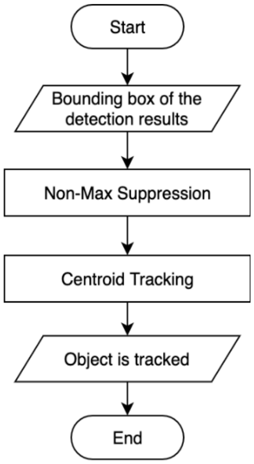

Fig. 4. Block Diagram of Tracking Object Process.

G. *Implementation*

Following the completion of the testing stages and the development of a good model for detecting human class objects, in this case, the instructor, the implementation is carried out in the classroom.

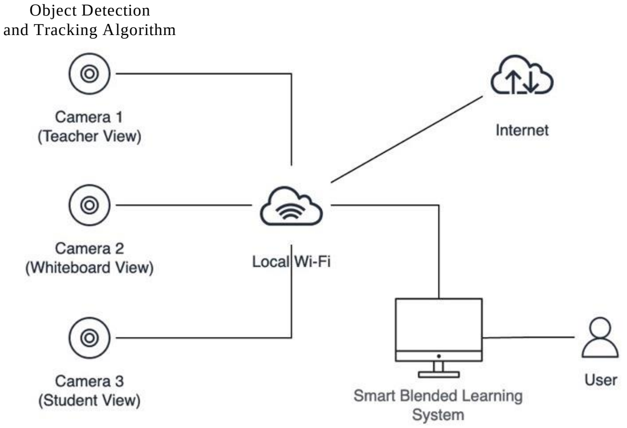

Fig. 5. The Architecture of the Smart Blended Learning System.

The MobileNet SSD object detection algorithm and centroid tracking are only applied to camera 1, pointing to the teacher, as shown in Fig. 5. The algorithm will monitor and zoom in on the teacher, ensuring that the video results in the video conference constantly center on the teacher.

IV. Result And Discussion

The F1 score test is used to evaluate the performance of the object detection algorithm by taking precision and recall into account. In 5,079 steps, the human object class is trained according to the results in Fig. 6. The resulting mAP value is obtained by employing transfer learning techniques, which allow the trained model to converge faster because the initialization of weights in the convolution process does not begin at random but rather uses the weights value of the model that has been trained on a large dataset with a variety of objects. This transfer learning technique enables the model to understand the pattern of each object class being taught more quickly.

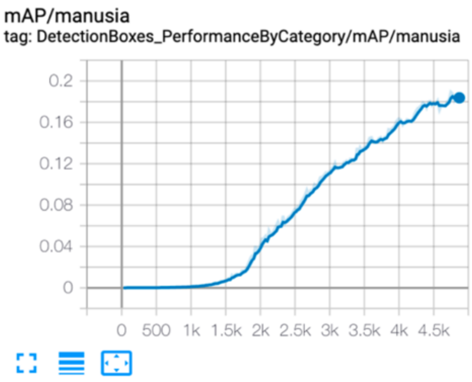

Fig. 6. Mean Average Precision.

The transfer learning technique allows the trained model to converge faster since the initialization of weights in the convolution process does not begin at random but instead employs the weights values of the previously trained model on big datasets with various objects. This transfer learning strategy enables the model to recognize the pattern of the object class being learned more quickly. Fig. 7 depicts the statistics of the loss value produced in each iteration.

Because this study focuses on predicting the positive class rather than the negative class, average precision and recall data are employed; suppose we have a total of 10,000 pieces of data. However, only 40 data points have positive labels, while 9,960 data are classified as negative. By predicting all negative classifications, the algorithm can achieve an accuracy of more than 99 percent with this data composition. The resulting model is less efficient if it can only be accurate in the negative class. Metric precision and recollection are required to address this difficulty. The mAP value varies from 0 to 1, with 0 being the lowest and 1 being the highest. The trained model's final mAP value on box detection performance for the human class is 0.1926.

The loss value is calculated from step to step till the training procedure is terminated. Losses are calculated in three ways: classification loss, localization loss, and regularization loss. Classification loss displays the error value for the object class's classification results. Localization loss is a number that expresses the error in determining the position of an object's bounding box during the inference process. Meanwhile, regularization loss is a process that adjusts or reduces the coefficient to zero, which is vital for assisting the trained network is converging more quickly. Based on Fig. 7, the last classification loss value is 6.02, the last localization loss value is 2.15, the last regularization loss value is 0.346, and if they are added together, the total loss is 8.516. The smaller the loss value, the better the performance of the resulting model.

Furthermore, the model's capacity to detect distance and camera height is tested to see how well the model performs, which will subsequently be utilized to detect things for teachers with varied detection distances and camera installation heights. The distances tested in this test are 2 meters, 4 meters, 6 meters, 8 meters, and 10 meters. While the camera tested has a height of 1.5 meters and 3 meters. Fig. 8 depicts an example of a test with varying distances.

The model testing results with varied distances and camera heights are shown in Table II. The dataset utilized was gathered independently by photographing human things at various distances and heights. The test results utilizing a camera location at the height of 1.5 meters revealed that the model could recognize human objects at detection distances of 2, 4, and 6 meters, as indicated by an F1 score of 1. It was discovered that the model's capacity to identify human objects is good at distances of 2, 4, and 6 meters while at a camera height of 3 meters. Object identification performance decreased significantly at a distance of 10 meters with an F1 score of 0.28.

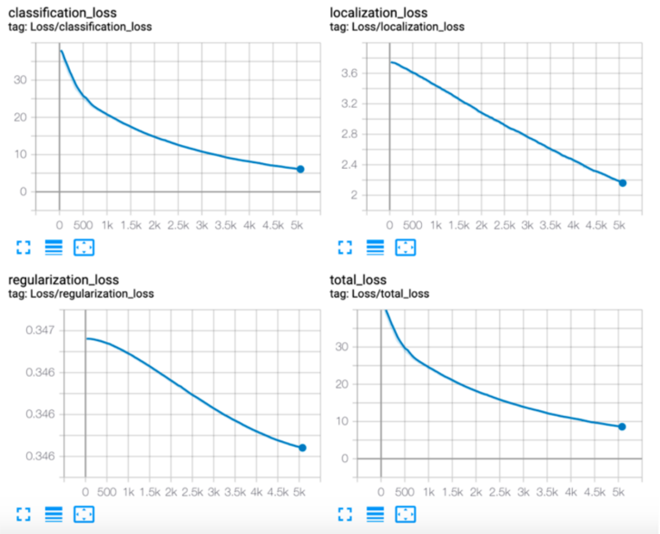

Fig. 7. Statistics and Loss Function Values.

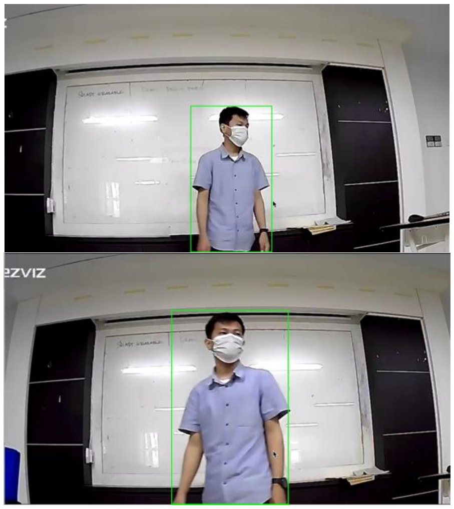

Fig. 8. Detection Results at Different Distances.

TABLE II. F1-MEASURE FOR DIFFERENT DISTANCES AND HEIGHTS OF CAMERA

No	Object Class	Height	Distance				
			2 m	*4 m*	*6 m*	*8 m*	*10 m*
1	Person	***1.5 m***	1	1	1	0.75	0.75
		3 m	0.92	1	0.8	0.76	0.28

This study adds new parameters to the detector network layer, specifically a new feature map with a depth of 64 at the network layer's end. The feature map is made smaller than the previous layer, which has a depth of 128. The addition of this layer attempts to provide the trained model the ability to distinguish smaller or farther away objects. The test results following the addition revealed that the detection results were steadier from a distance of 2 meters to 6 meters. However, at a distance of 8 and 10 meters, the accuracy value fell dramatically. Measurement results with varying distances are carried out in a room with sufficient light intensity, and limited lighting sources can affect the results of object detection in the system.

This study success performs object detection and tracking technology to shoot a wide area in the classroom so that wherever the teacher moves, the video will still focus on the teacher in the center of the video. It will provide space for teachers to teach and move freely in the classroom.

V. Conclusion and Future Work

We created a learning system solution in the form of a smart blended learning system based on a video camera in this framework by adding object detection and tracking. This solution has made online and offline learning more participatory for teachers and students.

So far, we've successfully implemented and piloted a smart blended learning system in a real-world classroom setting. A static camera is utilized in this video. Our next main project will involve the usage of a dynamic camera with a pan and tilt mechanism.

The design and implementation of our smart blended learning system framework are still ongoing, emphasizing the difficulties that continue to emerge and must be addressed to achieve a smart blended learning system framework in a global formulation.

Acknowledgment

The authors are grateful to the Indonesian Ministry of Education, Culture, Research, and Technology for funding this research through the Higher Education Leading Applied Research (Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi or PTUPT) scheme in 2021. We are obliged to Universitas Negeri Makassar (UNM) to provide the facilities and equipment needed for this research to be carried out.

References

[1] A. D. Ekawati, L. Sugandi, and D. L. Kusumastuti, "Blended learning in higher education: Does gender influence the student satisfaction on blended learning?," Proc. 2017 Int. Conf. Inf. Manag. Technol. ICIMTech 2017, vol. 2018-Janua, no. November, pp. 160–164, 2018, DOI: 10.1109/ICIMTech.2017.8273530.

[2] A. Al-Hunaiyyan and S. Al-Sharhan, "The design of multimedia blended e-learning systems: Cultural considerations," 3rd Int. Conf. Signals, Circuits Syst. SCS 2009, pp. 1–5, 2009, DOI: 10.1109/ICSCS.2009.5412342.

[3] B. Zhang, "Computer vision vs. human vision," in 9th IEEE International Conference on Cognitive Informatics (ICCI'10), 2010, p. 3. DOI: 10.1109/COGINF.2010.5599750.

[4] A. Wahid, S. Luhriyani, Nurhikmah, J. M. Parenreng, M. F. B, and M. I. Nur, "Smart Campus Framework: A Solution for New Normal Education System," in 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2021, pp. 266–271. DOI: 10.1109/ICITISEE53823.2021.9655952.

[5] J. Yang, H. Yu, and N.-S. Chen, "Using blended synchronous classroom approach to promote learning performance in rural area," Comput. Educ., vol. 141, p. 103619, Jul. 2019, DOI: 10.1016/j.compedu.2019.103619.

[6] M. Hastiea, I. C. Hung, N. S. Chen, and Kinshuk, "A blended synchronous learning model for educational international collaboration," https://doi.org/10.1080/14703290903525812, vol. 47, no. 1, pp. 9–24, Feb. 2010, DOI: 10.1080/14703290903525812.

[7] J. Carman, "Blended learning design: Five key ingredients," p. 11, Jan. 2005.

[8] M. Abisado, "A Flexible Learning Framework Implementing Asynchronous Course Delivery for Philippine Local Colleges and Universities," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, pp. 413–421, Jun. 2020, DOI: 10.30534/ijatcse/2020/6591.32020.

[9] Q. Wang, C. L. Quek, and X. Hu, "Designing and Improving a Blended Synchronous Learning Environment: An Educational Design Research," Int. Rev. Res. Open Distrib. Learn., vol. 18, no. 3, pp. 99–118, May 2017, DOI: 10.19173/IRRODL.V18I3.3034.

[10] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," ACM Comput. Surv., vol. 38, no. 4, 2006, DOI: 10.1145/1177352.1177355.

[11] B. Zhong et al., "Visual tracking via weakly supervised learning from multiple imperfect oracles," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., no. July 2010, pp. 1323–1330, 2010, DOI: 10.1109/CVPR.2010.5539816.

[12] A. Salhi and A. Y. Jammoussi, "Object tracking system using Camshift , Meanshift and Kalman filter," World Acad. Sci. Eng. Technol., vol. 6, no. 4, pp. 674–679, 2012.

[13] Z. Rahman, A. M. Ami, and M. A. Ullah, "A Real-Time Wrong-Way Vehicle Detection Based on YOLO and Centroid Tracking," 2020 IEEE Reg. 10 Symp. TENSYMP 2020, no. June, pp. 916–920, 2020, DOI: 10.1109/TENSYMP50017.2020.9230463.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 580–587, 2014, DOI: 10.1109/CVPR.2014.81.

[15] R. Girshick, "Fast R-CNN," Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, pp. 1440–1448, 2015, DOI: 10.1109/ICCV.2015.169.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017, DOI: 10.1109/TPAMI.2016.2577031.

[17] Y. Li, H. Huang, Q. Xie, L. Yao, and Q. Chen, "Research on a Surface Defect Detection Algorithm Based on MobileNet-SSD," Appl. Sci., vol. 8, no. 9, p. 1678, Mar. 2018, DOI: 10.3390/app8091678.

[18] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv1704.04861 [cs], Feb. 2017, [Online]. Available: http://arxiv.org/abs/1704.04861.

[19] A. Alimudin and A. F. Muhammad, "Online video conference system using WebRTC technology for distance learning support," Int. Electron. Symp. Knowl. Creat. Intell. Comput. IES-KCIC 2018 - Proc., pp. 384–387, 2019, doi: 10.1109/KCIC.2018.8628568.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," arXiv, 2018.

[21] W. Liu et al., "SSD: Single Shot MultiBox Detector," arXiv1512.02325 [cs], vol. 9905, pp. 21–37, Jan. 2016, doi: 10.1007/978-3-319-46448-0_2.

[22] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS -- Improving Object Detection With One Line of Code," arXiv1704.04503 [cs], Jun. 2017, [Online]. Available: http://arxiv.org/abs/1704.04503.