Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.041030
Article

SlowFast Based Real-Time Human Motion Recognition with Action Localization

Gyu-Il Kim[1], Hyun Yoo[2] and Kyungyong Chung[3,*]

[1]Department of Computer Science, Kyonggi University, Suwon, 16227, Korea
[2]Contents Convergence Software Research Institute, Kyonggi University, Suwon, 16227, Korea
[3]Division of AI Computer Science and Engineering, Kyonggi University, Suwon, 16227, Korea
*Corresponding Author: Kyungyong Chung. Email: dragonhci@gmail.com
Received: 08 April 2023; Accepted: 23 May 2023; Published: 28 July 2023

Abstract: Artificial intelligence is increasingly being applied in the field of video analysis, particularly in the area of public safety where video surveillance equipment such as closed-circuit television (CCTV) is used and automated analysis of video information is required. However, various issues such as data size limitations and low processing speeds make real-time extraction of video data challenging. Video analysis technology applies object classification, detection, and relationship analysis to continuous 2D frame data, and the various meanings within the video are thus analyzed based on the extracted basic data. Motion recognition is key in this analysis. Motion recognition is a challenging field that analyzes human body movements, requiring the interpretation of complex movements of human joints and the relationships between various objects. The deep learning-based human skeleton detection algorithm is a representative motion recognition algorithm. Recently, motion analysis models such as the SlowFast network algorithm, have also been developed with excellent performance. However, these models do not operate properly in most wide-angle video environments outdoors, displaying low response speed, as expected from motion classification extraction in environments associated with high-resolution images. The proposed method achieves high level of extraction and accuracy by improving SlowFast's input data preprocessing and data structure methods. The input data are preprocessed through object tracking and background removal using YOLO and DeepSORT. A higher performance than that of a single model is achieved by improving the existing SlowFast's data structure into a frame unit structure. Based on the confusion matrix, accuracies of 70.16% and 70.74% were obtained for the existing SlowFast and proposed model, respectively, indicating a 0.58% increase in accuracy. Comparing detection, based on behavioral classification, the existing SlowFast detected 2,341,164 cases, whereas the proposed model detected 3,119,323 cases, which is an increase of 33.23%.

Keywords: Artificial intelligence; convolutional neural network; video analysis; human action recognition; skeleton extraction

1 Introduction

Recently, the field of image analysis has progressed rapidly as it has been combined with artificial intelligence (AI). Automated image analysis is required in the social safety field where image surveillance equipment such as closed-circuit television (CCTV) is used. To prevent and respond to disasters, accidents, and crime in real time, it is necessary to analyze diverse types of information included in images, such as situation prediction, human action analysis, and face recognition. However, limited data size and low performance speed preclude current image data analysis processes from performing real-time extraction [1].

Image analysis technology uses two-dimensional consecutive-frame data obtained from image sensors such as CCTVs and web cameras as raw data [2]. Subsequently, object classification, detection, and associative relations are applied to the extracted raw data, to analyze the diverse meanings within images [3]. In practice, AI-based image analysis studies using CCTVs are associated with social safety, patrolling, national defense, and traffic, with commercialization in specialized fields such as traffic control and vehicle license plate recognition [4]. However, universally, the use of image data analysis in practical environments is limited due to the low accuracy of image analysis systems. A universal image analysis system is required to extract and classify objects and analyze the dynamic relation between objects. In particular, classification of human motions and diverse types of objects is required to analyze associative relations such as movements and contacts. As such, situation detection technology captures parts based on dynamic situations such as motion, movement, and contact between people within an image, interpreting the part differently depending on the external environment and situation, by appropriately applying each factor [5]. In particular, the motion recognition field requires a high level of interpretation for the analysis and understanding of the associative relation between complicated human joint motions and diverse objects. In addition, as image data size is comparatively large, the performance of the response is low. One of the representative motion recognition algorithms is the deep learning-based human skeleton detection algorithm [6]. The skeleton detection algorithm was studied to provide a fundamental solution to human posture estimation. However, the fundamental problem of having to capture all the human joints within an image, slows down the performance, necessitating an additional algorithm for motion analysis. Recently, the SlowFast network algorithm [7], demonstrated comparatively fast and excellent performance in motion analysis, using images focused on by the camera and a small number of adjacent human bodies in all stages of learning, classification, and evaluation. However, in real-world industrial environments, the cameras often have a wide field of view and long visibility range, making it difficult to overcome the practical issues involved in accuracy improvement solely through enhancements in the internal structure of SlowFast. Therefore, even with various improved forms of SlowFast that have been recently researched and developed, low accuracy still persists. We attribute these to the fundamental problems with the algorithmic structure of SlowFast, which centers around only a few main subjects. Specifically, when dealing with wide backgrounds and multiple subjects, each performing different actions, the quality of results further deteriorates. Therefore, an algorithm capable of removing unnecessary backgrounds and preprocessing data based on core human motions is necessary.

In this study, a three-layer model that combines three models is proposed to address the existing issues of response time and accuracy. The three-layer model comprises you only look once (YOLO) model [8] for detecting the main subject of the behavior, a DeepSORT model [9] for tracking the movement of the main subject, and an internally improved SlowFast model for extracting real-time action detection results. The YOLO and DeepSORT models are used for preprocessing to remove the unnecessary background objects excluding the main subject. The three-layer model produces more

accurate results with the internally improved SlowFast model exhibiting fast response performance. Combining these two types of deep learning algorithms provides a more realistic alternative to CCTV type video input environments.

The extracted amount, which refers to the percentage of detected frames, was evaluated to determine the extraction performance of the proposed model. In addition, the accuracy and speed of the response associated with action classification were evaluated. The contributions of the proposed method are as follows:

- The general SlowFast model uses wide-angle images and as such, accuracy may be reduced because of unnecessary background. The one-shot object detection model and sort-based object tracking algorithm eliminate unnecessary background and improve performance.
- A typical SlowFast model processes data in 64 frames. Consequently, the response speed is reduced. The proposed center-frame structure improves internal processing procedure of SlowFast on a frame-by-frame basis as well as response speed.
- The response speed of the configured model is close to that of a real-time response. Therefore, it is possible to monitor and detect situations in real time, which has applicability to various monitoring fields.
- High accuracy and near-real-time response speeds reduce inefficiency and cost of traditional manual image monitoring surveillance systems, thereby enabling proactive action against a variety of hazardous situations.

The remainder of this paper is organized as follows. In Section 2, current object detection and human motion recognition methods are described along with motion recognition technology trends. Section 3 describes the proposed SlowFast-based real-time human motion recognition method, which employs object detection and tracking. The results and performance evaluation are described in Section 4, and conclusions are presented in Section 5.

2 Related Work

2.1 Object Detection Techniques Based on Deep Learning

Pixel characteristic tracking and trajectory classification have been used in studies aiming at extracting meaning from 2D RGB images, to understand the motion within images. Most previous studies have focused on methods based on optical flow [10]. Pixel characteristic tracking determines or predicts the movement of objects by calculating the pixel direction and distribution between frames, and thus analyzes the movement patterns of objects within an image. Some representative algorithms include the Luca–Kanade [11] and Farneback [12] algorithms. The Luca–Kanade Algorithm accumulates pixel windows (3×3) in the form of a pyramid and determines the movements of objects through similar pixel movements in the pixel windows. The Gunnar–Farneback algorithm detects pixel intensity changes between frames and converts them into vectors. This process provides object direction. However, these methods are vulnerable to noise, such as those generated by body motions and situations associated with rain or wind. To overcome this problem, supplemental efforts, which involve the installation of diverse sensors and multiple cameras, are underway. However, methods based on simple algorithms have several drawbacks, such as increased cost and design complexity arising from the number of equipment installations. Approaches such as increasing the number of sensors and installing multiple cameras, may resolve such problems at the expense of increased cost [13,14].

Recently, convolutional neural network (CNN)-based object detection technology has been applied, and region-based CNN (RCNN) [15], two-shot object detection [16], and one-shot object detection [17] methods such as YOLO are being developed. Two-shot object detection requires two stages to perform object detection, whereas one-shot object detection requires only one stage. Two-shot object detection consists of two stages: extraction of regions of interest (ROIs) and their classification. ROI extraction refers to the process of finding a region in which objects may exist. Although high performance can be expected from such a process, the response speed is limited because of increased computational complexity. In contrast, one-shot object detection extracts and classifies regions of interest in one stage. Therefore, computational complexity is reduced resulting in a faster response speed than two-shot object detection; however, the performance is inferior to that of two-shot object detection.

Deep learning (DL) is a powerful machine-learning technique that achieves state-of-the-art results in various tasks such as image classification [18], object detection [19], natural language processing [20], and speech recognition [21]. In recent years, DL has become increasingly popular because, unlike traditional machine learning methods, it has the ability to learn more complex patterns in large datasets. DL has been used to solve various problems in different fields such as climate change prediction and environmental analysis. Therefore, DL is a powerful tool that has the potential to address many of the world's most pressing problems.

2.2 Human Motion Recognition Techniques

In general, recent human motion analysis methods are based on human skeleton detection using deep learning. Human skeleton detection extracts the coordinates of each body part by using an algorithm based on a convolutional network. The motions are then analyzed by connecting each joint object. Therefore, it is possible to classify human motions and actions using an algorithm. To achieve this, it is necessary to extract the background and human objects separately from an image, and an additional algorithm is required for practical motion and situational judgment.

Pose estimation detects keypoints and continuously tracks their relationships. Pose estimation can be divided into two methods: bottom-up and top-down. The top-down method first detects humans (i.e., objects). This method estimates poses within a bounding box in which humans are detected. As poses are estimated after detecting humans, the method is limited in that there are a number of stages and estimation of poses cannot be completed when no humans are detected [22]. By contrast, the bottom-up method first detects the keypoints and estimates human poses by connecting the detected keypoints. This method is also limited because there are a large number of keypoint connection combinations, and it takes a long time to determine the appropriate combinations [23]. However, because this method does not require that humans be detected first, it is convenient to apply to real-time systems. Fig. 1 shows the differences between top-down and bottom-up methods.

The pose estimation field is largely divided into two approaches: single-stage and multistage. The single-stage approach focuses primarily on the design of a basic network structure. The hourglass approach, which is the most typical single-stage approach, considers information from other scales during upsampling. Cascaded pyramid network (CPN), another single-stage structure, has the strengths of both CPN and hourglass [24]. Multistage approaches focus on the design of multiple structures to enhance performance. A convolutional pose machine (CPM) has a multistructure consisting of several convolutional and pooling layers. The stacked hourglass structure has several stacked hourglass layers [25], as suggested by the name. In addition, there are diverse models that

utilize multi-context attention, such as ResNet. In general, the multi-stage approach shows better performance than the single-stage approach, but each stage requires a sophisticated design.

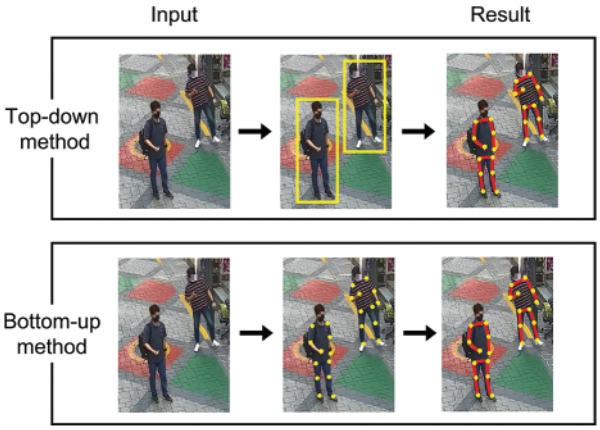

Figure 1: Method of top-down and bottom-up

The high-resolution network (HRNet) is a representative pose estimation algorithm [26]. HRNet expresses joints as coordinates in an image using a heat map. The HRNet model is improved by maintaining high-resolution expressions during deep learning. This method begins with a high-resolution network and adds one high-resolution subnetwork at a time to develop additional stages. It then connects the multi-resolution subnetworks in parallel, thereby performing repeated multiscale fusions such that each high-resolution expression repeatedly receives and combines information from other parallel expressions. Because this method repeats multiscale fusions, unlike pre-existing similar models, it does not require intermediate heat map control. Based on the final output of the network obtained through this process, a heat map capable of predicting the joint location is extracted more accurately. The HRNet model performs better than pre-existing methods in detecting key points and estimating multiple poses when applied to common objects in context (COCO) and Max Planck Institute information (MPII) datasets [27]. Excellent results were obtained using the PostTrack dataset [28]. However, although HRNet is an excellent model, since all the heat maps have to be calculated, its low speed is a limitation.

2.3 Motion Recognition Technology Trends

In a recent study on video recognition, human actions were recognized and classified by configuring an artificial neural network (ANN) based on deep learning. Two-stream neural networks [29] and human-skeleton detection [30] are representative methods. A representative SlowFast network model with two-stream networks aims to detect diverse situations and actions using data consisting of consecutive 2D images. SlowFast consists of two-stream neural networks that do not use the optical flow method of previous studies, was the announcement by Facebook artificial intelligence research (FAIR) [31]. SlowFast collects, connects, and learns several associated back-and-forth frames. The SlowFast model internally combines two algorithms, slow and fast, and configures a two-stream network model. The slow model analyzes the overall environment and situation of an image, whereas the fast model captures dynamic movements. The slow model consists of a 3D convolutional network, similar to a previously developed action recognition model. The slow model classifies images by entering the received data into a convolutional network as a video clip in which the frames overlap as time progresses. This model analyzes spatial situations based on overall content. The fast model

is designed to capture rapidly changing motions from consecutive frames. This model detects pixels that change quickly as time progresses using a high frame rate in consecutive images. The fast model decreases the number of image channels, thereby reducing the overall computational cost. The lateral connections are used to combine the results. The two-stream network methods display high accuracy and excellent results, and placed first in the action field of the AVA challenge [32]. In addition, SlowFast internally configures an action recognition deep learning model based on an artificial neural network using a convolutional network. This model collects each frame, creates a 3D spatiotemporal dataset, and analyzes the created dataset using a 3D convolutional network. However, the preexisting SlowFast model is limited because of low response speed in data processing.

Ramakrishna et al. [33] proposed a pose estimation algorithm, utilizing a new method that synthesizes several classifiers and learns the motion inference process directly, thereby exhibiting better performance than preexisting pose estimation. However, because computational complexity is high, and it is difficult to estimate occlusion poses, this model is difficult to apply to real-time systems. To resolve this problem, Wei et al. [34] proposed a pose-estimation algorithm in which a convolutional network is employed. This method applies a convolutional network in lieu of the group of classifiers proposed by Ramakrishna et al. [33], thereby exhibiting better performance and higher occlusion pose detection. However, the algorithm has limitations in that it is difficult to estimate poses when there are many people. Therefore, it is necessary to develop an algorithm that can estimate the poses of multiple individuals with high response speeds. Yoo et al. [5] proposed a deep-learning action classification model based on a skeleton analysis algorithm. This model is a lightweight skeleton pattern algorithm that classifies and recognizes human actions in real time. They designed 3D vectors using skeletal coordinates and classified images using a DNN model. The proposed method has high accessibility and low computational cost. However, the computational cost increases with the number of objects in the image, resulting in a slow response. In addition, recent object detection models have limitations in terms of computational complexity because of the relatively large training data. The low response speed is also a limitation in real-time object detection. To overcome this limitation, a model with high response speed is required. Arunnehru et al. [35] utilize the spatio-temporal interest points (STIPs) approach along with Two Dimensional-Difference Intensity Distance Group Pattern (2D-DIDGP) and Three Dimensional-Difference Intensity Distance Group Pattern (3D-DIDGP) to construct unique and discriminative feature description methods for enhancing action recognition rates. Arunnehru et al. [35] demonstrate excellent performance in action recognition using the UT-Interaction dataset with Support Vector Machines (SVM) and Random Forest (RF) classifiers. However, the UT-Interaction dataset differs from real outdoor CCTV environments with wide-angle focus due to issues of resolution and object sizes within the video. Therefore, there is a need for algorithms that perform well on datasets that resemble real-world environments.

3 SlowFast Based Real-Time Human Motion Recognition with Action Localization

Most motion analysis algorithms based on two-stream networks are designed and trained to operate with a limited number of key persons during learning, classification, and evaluation. Therefore, these algorithms track human motion, which is a key factor associated with events in images. Through this process, unnecessary background is removed from images, and using consecutive images of the tracked persons as input to the two-stream network, it is possible to improve performance. To realize such procedures in real time, it is necessary to improve and combine heterogeneous object detection, tracking, and motion analysis algorithms that differ in direction according to the field of study. The configured SlowFast-based real-time human motion recognition system with action localization is shown in Fig. 2.

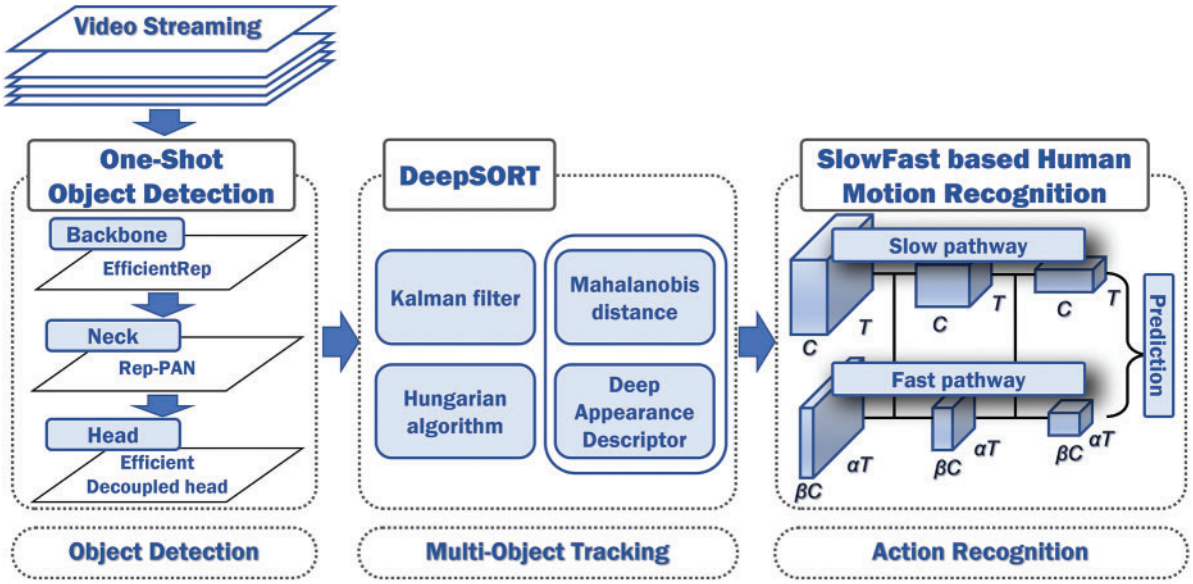

Figure 2: Progress of SlowFast based real-time human motion recognition with action localization

The configured model consisted of three models for action extraction. In the first stage, as shown in Fig. 2, a one-shot object detection model is used to collect the location and size of an object. In the second stage, the collected data are used as input for the DeepSort algorithm to determine whether the same object exists in consecutive frames. In the third stage, the images of an object recognized as being the same object by the DeepSort algorithm, are separated and converted into a consecutive image structure, whereby the converted structure is entered into the SlowFast algorithm. Therefore, the SlowFast algorithm operates more effectively by analyzing consecutive image videos separated from one object. Using such complex deep-learning algorithms, unnecessary background can be removed from human objects within an image. Subsequently, consecutive video images separated from one or two human objects are analyzed using a motion-analysis algorithm based on two-stream networks, generating a divide-and-conquer effect, thereby achieving higher accuracy and response performance.

3.1 Object Detection and Tracking Combining One-Shot Object Detection and DeepSort

Improved motion-classification performance can be achieved by minimizing unnecessary backgrounds. Therefore, for background removal, object detection was first performed, using YOLO v6 one-shot object-detection models with high response speeds, followed by sort-based object tracking.

Fig. 3 shows the internal configuration of the one-shot object detection method. The one-shot object detection model used in this study consisted of a backbone, neck, and head. Reparameterization is the key to YOLO v6. The EfficientRep backbone used in this study consisted of 53 convolutional layers. The RepBlock in the backbone uses RepVGG during training with an additional ReLU. In the neck region, Rep-PAN, which comprises upsampling and concentration, extracts features. The head learns the classification and localization, which are decoupled from each other. The head, which consists of three detection layers, proceeds with boundary box creation and class prediction based on the characteristics of the neck. When configuring a simplified neural network system where speed is important, it is necessary to effectively configure the system considering the computing power setting to prevent a decrease in accuracy.

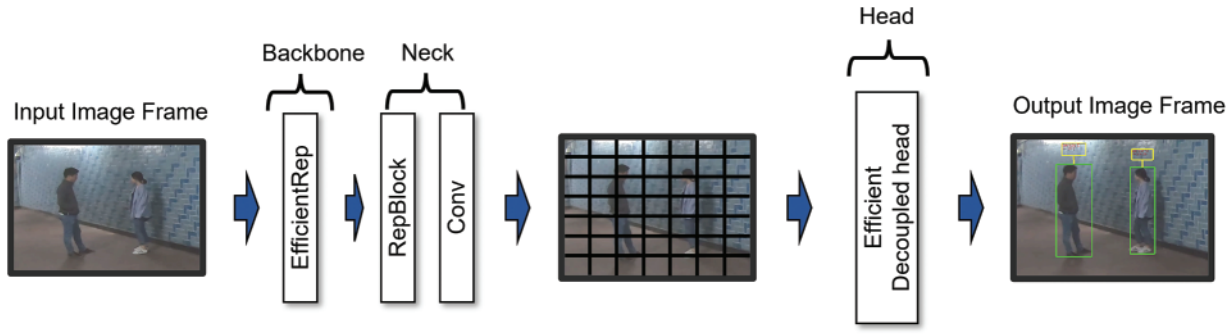

Figure 3: Structure of one-shot object detection

The COCO dataset was used for data learning in the configured one-shot object detection. The COCO dataset contains information such as object detection, segmentation, and captions. It contains more than 300 K images and 200 K labels, and the labels consist of 80 object categories. As this dataset contains a large amount of image data, it is suitable for human detection. The detected data of individuals are used as input for the DeepSort algorithm where they undergo a process to determine whether the same object exists in consecutive frames. Fig. 4 shows the results of one-shot object detection. In Fig. 4, the original image is on the left, and the normalized results obtained from the transfer of the detected object to a separate memory are on the right.

Figure 4: Output of one-shot object detection

DeepSort comprises a Kalman filter and the Hungarian algorithm, and uses the data created by the detector algorithm. The Kalman filter is used to process the noise generated during detection, whereby the status prediction and measurement update stages are alternated according to the input of each frame. In the status-prediction stage, the probability distribution calculated during the previous measurement update is used to predict the current status distribution, whereas in the measurement-update stage, the predicted and actual observed probability distributions are used to update the posterior probability. Consequently, tracking an image resolves the linearity problem, whereby the object instantaneously disappears or does not appear. The Hungarian algorithm searches for optimal matching and identifies objects using the assignment problem-solving method, which is formulated as a problem of minimizing the cost incurred while traveling between peaks on two graphs. The Mahalanobis distance quantifies the standard deviation from a particular mean value along with

the probability that a value will occur as well as the ideal value [36]. The structure of the DeepSort algorithm is shown in Fig. 5.

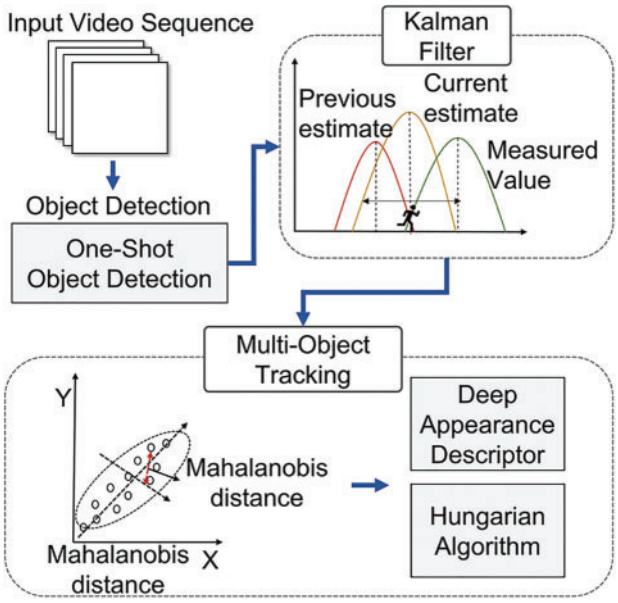

Figure 5: Structure of DeepSort

In the final stage, object images recognized by the DeepSort algorithm as the same object are separated and converted into a structure of consecutive image frames, which are then entered into the SlowFast algorithm. The form of the entered image is shown in Fig. 6 illustrating how the object detected by the one-shot object detection model is tracked using DeepSort. The right side of Fig. 6 shows the tracking process. Each number represents the frame order.

Therefore, it is necessary to correct the SlowFast network. Because the SlowFast network is a heavy algorithm, its use in tracking multiple objects simultaneously in real time is limited. Therefore, it is necessary to perform separate clipping of key persons and objects tracked in an image prior to entering them into the SlowFast Network. If there is only one person in an image, it can be entered directly into the SlowFast Network. However, in the case of more than two people, the image must be analyzed based on the persons involved in a critical incident. In this study, the criterion for the occurrence of important events was two-person contact. Because such contact between two persons inevitably occurs in most crime situations, such as assault, kidnapping, and theft, this standard is a realistic option in situations with limited computing power. When more than two persons exist in an image, the two persons who made the most intimate contact were entered into SlowFast, and the results were extracted based on the entered data.

As the SlowFast network learning dataset contains an inner margin, a separate image inner margin is required. The inner margins represent background borders of the dataset. Kinetics 400 was used as the SlowFast network learning dataset [37]. The data content of the Kinetics 400 is learned with a focus on people with some background, having realistic forms, for applicability to real-life scenarios. Therefore, for images extracted from a video, the outline of an object should not be used as a standard; instead, a separate inner margin should be set. It is necessary to adjust the size of the inner margin according to the learning and evaluation data. An inner margin of 10% was used in this study. In addition, the one-shot object detection model uses the YOLO model. YOLO v6, one of the most

standardized YOLO versions that enables convenient performance comparisons, was used. YOLO divides an input image into multiple grid forms. It then creates bounding boxes for the objects in each section and generates a probability for detecting objects in each region, whereby these probabilities are used to improve speed and accuracy. YOLO requires comparatively simple processing and is suitable for real-time data analysis.

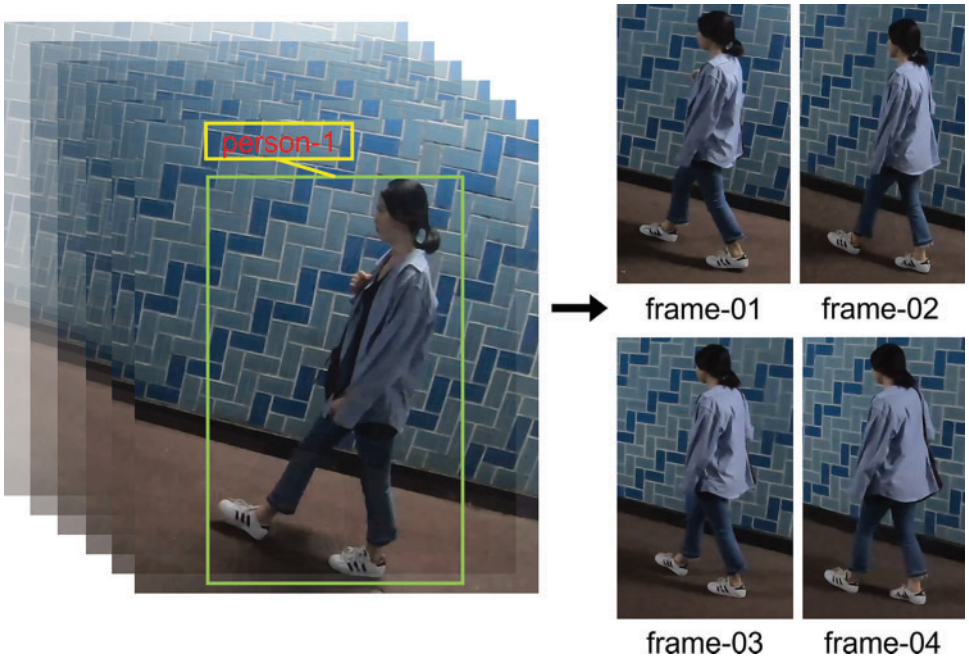

Figure 6: Output sample of DeepSort

3.2 Improved of SlowFast with Overlapping Frame Structure

For efficient video recognition, a lightweight model with a high accuracy is required. In this study, the configuration of the preexisting SlowFast network model was used as is, and the data processing structure was changed to improve the response performance using a GPGPU and configuring the data to have an overlapping frame structure. SlowFast internally combines a slow model for judging the surrounding environment with a fast model for motion classification to categorize a situation within an image and exhibits excellent accuracy.

Regarding the structure of the SlowFast network model used, the two-stream network structure combines two algorithms, known as slow and fast pathways, which were identically applied to both models. The slow pathway analyzes the overall environment and situation of an image, whereas the fast pathway captures dynamic movements. The internal structure of each pathway comprises a convolutional network. The fast pathway uses a high frame rate to detect pixel changes and decreases the number of image channels to reduce computational cost. The data structure of the fast pathway performs $\beta C \times \alpha$ T convolutions, where α is the speed ratio; β is the channel ratio; C denotes channel; and T is temporal length. The slow pathway consists of C $\times$ T convolutions. The values used for the internal configuration of SlowFast in this study were $\alpha = 8$ and $\beta = 1/8$. The structure and data flow of the configured SlowFast are shown in Fig. 7.

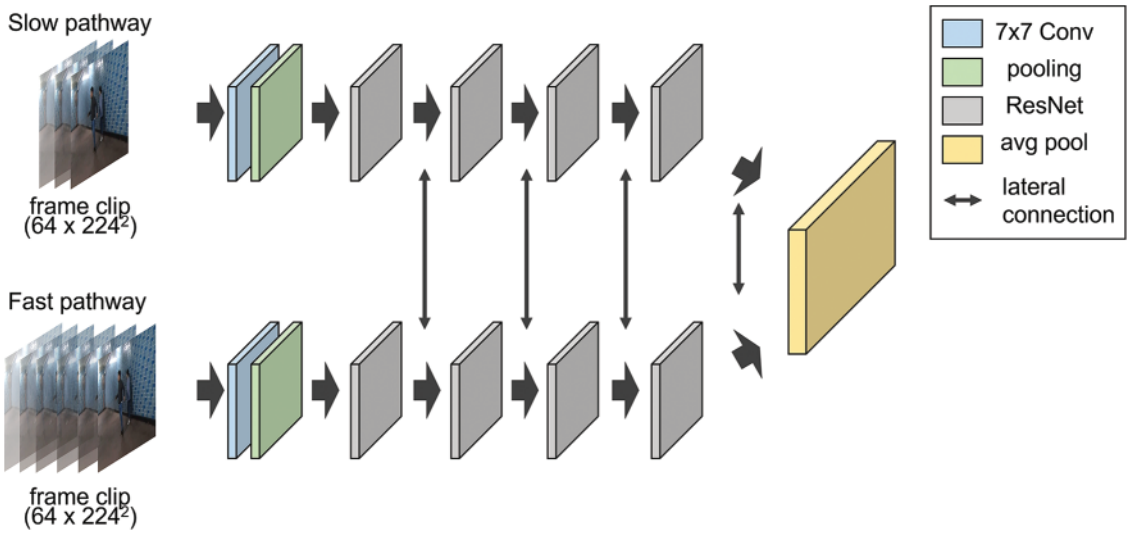

Figure 7: SlowFast structure and data flow

In Fig. 7, the content of a frame clip is represented by the number of frames, height, and width. The frame clip dimensions were 64 × 224 × 224 pixels, with conv, pool, and res denoting the convolution, pooling, and ResNet layers, respectively. Each slow and fast pathway consisted of one convolution layer and four ResNet blocks. The ResNet blocks had three, four, six, and three layers, respectively. Finally, an integrated global average pooling layer was created. The initial frame data pass through the convolution and pooling layers, and the slow pathway runs through a neural network similar to ResNet-50. As the fast pathway runs through a similar but smaller channel to classify images, it requires fewer computational resources, running through each RES module and combining mutual information through a lateral connection. Because such a configuration is better at detecting dynamic situations than a single slow model, a higher accuracy is obtained in analyzing data samples associated with body motions such as greeting and clapping. Using this structure, the SlowFast model can accurately classify images. However, the data input structure problem causes delays in obtaining initial analysis results. In addition, memory processing time increases as frames progress. Memory operation overhead and the critical event truncation phenomenon that occurs every time the image data structure is separated, result in slower speed and lower accuracy.

To resolve the problems of the SlowFast model, the frame data of the input image were improved into an overlapping structure to improve accuracy. The SlowFast network model is particular in that it accumulates and converts 64-frame unit image data. The image data were entered into a general SlowFast neural network with 64 × 224 × 224 pixel dimensions. An initial delay occurs until 64 frames are collected. Subsequently, the same frames are combined and evaluated, and this procedure is repeated. A delay occurs in deducing the results as the frames accumulate, incurring an overhead because of the removal and transfer of the frames accumulated in memory. In certain cases, the image in which a critical motion event occurred is separated and evaluated. Consequently, in practice, a time difference occurs between the actual observation of an image and the result, which leads to a few seconds of delay, and the process is repeated. A delay in the initialization process is inevitable because of the characteristics of the algorithm. However, it is possible to configure a model based on the first-in-first-out (FIFO) method by overlapping subsequent frames, which can be improved using general-purpose computing on graphics processing units (GPGPU). This operational method is known as

the overlapping frame method [38]. Configuring and repeatedly entering such a method into a neural network, results in higher accuracy. The overlapping frame method preprocesses memory operation units consisting of preexisting image units into frame units to decrease the overhead resulting from large-scale memory operations and introduces improvements such as performing numerous repeated operations using the GPGPU. It also converts data into frame units. The overlapping frame method is divided into improved preprocessing operations consisting of normalization and configuration of data forms into frame units. The operation of the proposed overlapping frame method is illustrated in Fig. 8.

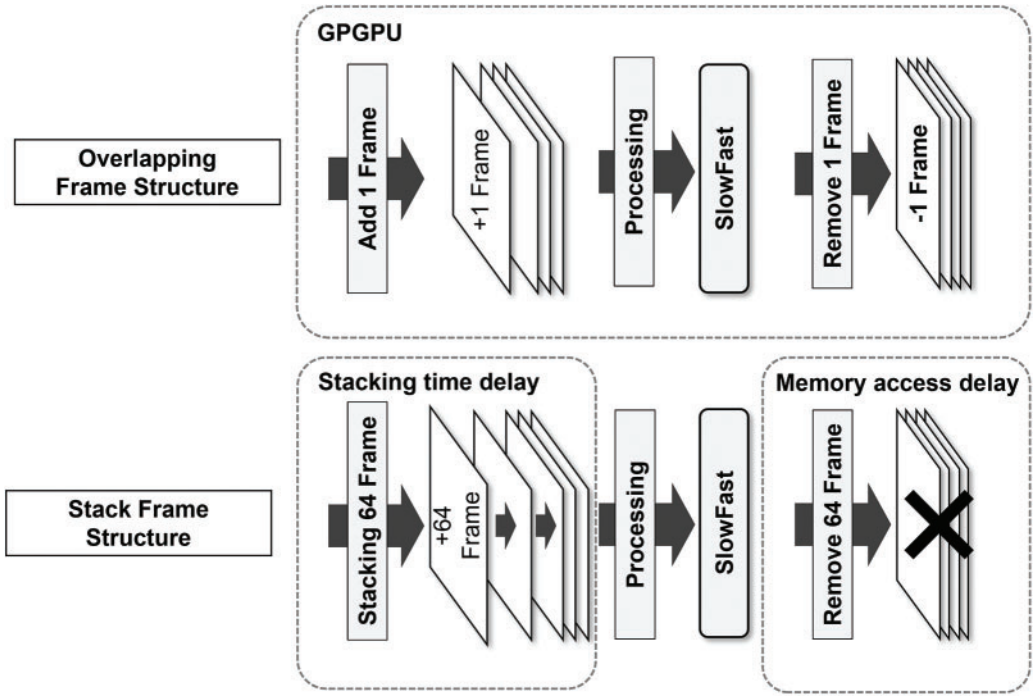

Figure 8: Progress in overlapping frame structure

In real-time image analysis, the overlapping frame structure consisted of accumulated 64 frames of initial size-adjusted images of 224 × 224 pixels. An additional frame was then attached to the back, and the oldest front-end frame was removed. Therefore, the overlapping frames had the same size of 64 × 224 × 224 pixels. This structure is similar to the FIFO structure, and because it does not require large-scale memory operations, the overhead decreases. The results were extracted by entering these data into the SlowFast network in frame units. This method is particularly useful because the classification results can be extracted per frame, and when T is greater than 64, more results can be extracted within the same amount of time. This increases the possibility of capturing the keypoints of a certain motion in a more general situation. Therefore, this method is effective in practical situations. In addition, in terms of practical use, the delay is reduced, as well as memory operations resulting from frame accumulation time, which are a disadvantage. SlowFast maintains its structure and advantages by optimizing the input frame's structure to enable real-time result extraction.

In the configured SlowFast, image data processing consists of learning and evaluation stages. During the learning stage, image data frames suitable for the overlapping frame structure are combined, and the data structure required for classification is created and maintained. Kinetics 400

was used for the learning and evaluation. These data consist of approximately 30,000 images classified into 400 types of human motion. During the learning stage, an accuracy evaluation was performed based on epochs. Fig. 9 shows the results obtained using the SlowFast model. SlowFast detects motions and displays actions against the detected motions.

Figure 9: Output of SlowFast

4 Result and Performance Evaluation

The accuracy, number of extracted frames, and response speed of the models were evaluated. Image data from videos were used for evaluation to confirm model accuracy in a more general image analysis setting. For the Kinetics 400 data, the distance between the camera and the subject for human motion is small, and because the camera follows the subject in most cases, it is difficult to evaluate general environments, as with CCTVs. AI Hub's human motion image data provided by the National Information Society Agency consist of images of people in videos captured using more general external cameras [39], whereby pose information is obtained through pose estimation. Therefore, we tested our model, which is specialized for analyzing wide-angle videos by removing the background, on this dataset. Of the diverse actions provided, seven classifiable motions (walking, stair climbing, direction-indicating push-ups, clapping, sit-ups, and crawling) were selected, and the accuracy of each motion was measured. Hardware consisting of an AMD Ryzen 7 5800X, 16 GB of memory, and an NVIDIA GeForce RTX 3080 Ti was used. Python version 3.9.7 and PyTorch version 1.12.0 were used to configure the neural network.

4.1 Confusion Matrix

The motion with the highest accuracy was evaluated to determine the accuracy of human motion image data from the AI Hub. The accuracy was evaluated using a confusion matrix, which is a matrix for comparing the predicted and actual values to measure training performance. This was a 2×2 matrix. The accuracy, precision, recall, and F1 scores were measured using the matrix elements. Fig. 10 shows the accuracy evaluation results obtained from the general SlowFast and proposed models.

Based on the evaluation results obtained using the confusion matrix, the accuracies of the OpenPose, Vanilla SlowFast, and proposed models were 70.08%, 70.16%, and 70.74 %, respectively. The accuracy of the proposed model is 0.58% higher than that of the baseline SlowFast model and 0.66% higher than that of the OpenPose model. This indicates that the improved model architecture is more effective for multiple evaluations.

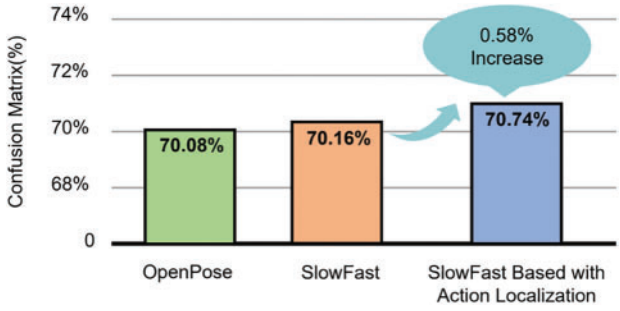

Figure 10: Confusion matrix result

4.2 Extraction Amount Evaluation

The amount of extraction was evaluated by comparing the number of detected frames to the total number of frames in the general model. Performance was measured by evaluating the extent to which the extraction amount of the proposed model increased compared to that of a general model. It is necessary to check whether objects are detected in all the frames in the video.

$$Extraction\ amount = \frac{Detection\ Frames}{Total\ Frames} \tag{1}$$

In Eq. (1), the total number of frames is the number of frames within the image, and Detection Frames are the number of frames that detect the target object within the image. In Eq. (1), if object detection is performed well, the resulting value is close to one. Fig. 11 shows the evaluation results of the extraction amount obtained from the general model, which combines SlowFast with DNN and the proposed model.

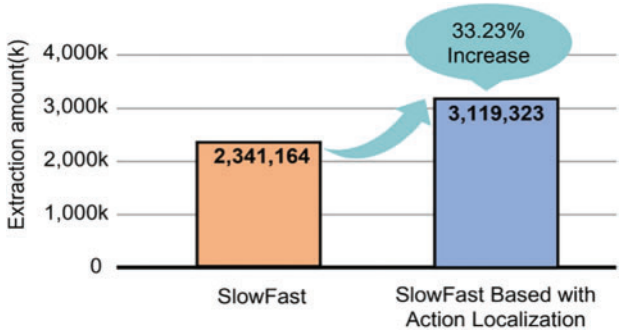

Figure 11: Extraction result

In Fig. 11, the total number of frames in the image was 17,911,868. The overall extraction amount increased by 33.23% from 2,341,164 to 3,119,323. These results indicate that the improved model tracks the detected objects separately and removes the background from the object images, thereby increasing the detection frequency of consecutive motions and the total detection amount. Fig. 12 shows the response speed evaluation results obtained from the existing SlowFast model, the SlowFast model to which the proposed structure improvement method is applied, and the proposed model. The existing SlowFast had a response time of 1.886 s. The improved SlowFast algorithm with an overlapping frame structure exhibited a response time of 0.046 s. This is an improvement, demonstrating that the response time decreased by 97.56%. This result shows that the improved SlowFast model preprocesses

the input data and improves the response time demonstrating that the effect of the proposed model is valid. However, because the final proposed model operates using several complex algorithms, the response is delayed. YOLO caused a delay of 0.007 s, whereas DeepSort caused a delay of 0.018 s. An additional delay of 0.025 s occurred compared to SlowFast, which has an overlapping frame structure. This demonstrates that the delay increased by 54.34%. Based on these results, it can be confirmed that accuracy increases; however, an increase in the computational cost accompanies. Therefore, the extraction amount increased by 33.23% compared to the preexisting single SlowFast model. In addition, the response time of the proposed model decreased by 97.56%. Finally, the response time of the model combining multiple algorithms decreased by 96.23%. Using this method, a near real-time classification can be performed. However, for commercial use, it is necessary to develop a more effective optimized configuration. These problems are expected to be addressed in future studies through hardware performance enhancements and algorithm improvements.

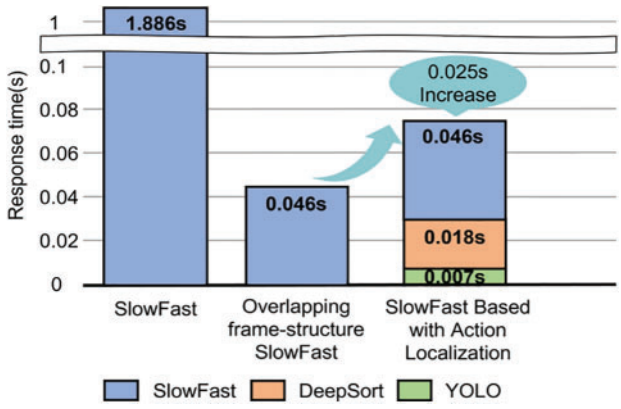

Figure 12: Response time result

4.3 Computational Complexity

The dataset used in this study consists of 12 actions with a total of 8,436 videos. The proposed model calculated the algorithmic complexity per frame. The original SlowFast model had a computation time of 1.886 s per frame, while our Overlapping frame-structure SlowFast model took 0.046 s per frame. The final model, which combined three models, took 0.071 s per frame. The average computation time for a single video in the original SlowFast model was approximately 4 h and 47 min, while our model took an average computation time of approximately 10 min and 49.65 s per video. Therefore, our model showed a performance improvement of approximately 96.23% in terms of computational complexity compared to the original SlowFast model.

5 Conclusion

An object detection and tracking model is combined with a human action analysis model based on deep learning, to obtain a model with improved accuracy and speed. This model can be incorporated into a video-based image surveillance system to promptly capture and respond to diverse situations. In this study, heterogeneous deep learning algorithms were combined, that is, SlowFast's data processing structure was combined with a frame structure to improve response speed and accuracy. The proposed model was not evaluated on Kinetics 400 data, but instead on AI Hub's abnormal action image data recorded by general external cameras and provided by the National Information Society Agency.

Frame extraction rate was 33.23% greater than that of the preexisting model, with near real-time response performance, making it possible to analyze the action patterns between persons or between a person and an object at high speed. In the actual industrial environment setting, where consecutive streaming data are entered, such structures exhibit a more enhanced effect. Therefore, it is possible to analyze images and detect particular abnormal actions in real time in CCTV-based image-control situations. An image control system based on the configured model can be utilized in diverse fields ranging from safety, such as fall and collision detection, to crime detection, such as assault and theft detection, to other fields, such as traffic accident and truce line intrusion detection. Therefore, the proposed model can introduce economic and industrial ripple effects in urban safety, police, national defense, and traffic control and is expected to play an important role in social safety monitoring systems. In this study, the extracted knowledge was applied to diverse analyses. In the future, we plan to apply data mining techniques to the data extracted in this study. Standardization of the extracted data is another possibility that can contribute to crime prevention. Therefore, it is necessary to develop a more effective and optimized configuration for commercial use. In particular, when multiple objects appear in the image, the computation load is extensive depending on the number of objects. Therefore, auxiliary techniques such as object limitation analysis are required for real applications. Thus, further research is required to make the algorithm lighter and simplify the data pipeline in future studies.

Acknowledgement: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A0 3040583). This work was supported by Kyonggi University's Graduate Research Assistantship 2023.

Funding Statement: The authors received no funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Díaz, J. B. Stephenson and M. A. Labrador, "Use of wearable sensor technology in gait, balance, and range of motion analysis," *Applied Sciences*, vol. 10, no. 1, pp. 234, 2019.

[2] K. Morimoto, A. Ardelean, M. Wu, A. C. Ulku, I. M. Antolovic *et al.,* "Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications," *Optica*, vol. 7, no. 4, pp. 346–354, 2020.

[3] K. Sun, B. Xiao, D. Liu and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. CVPR*, Long Beach, CA, USA, pp. 5693–5703, 2019.

[4] J. Kim and K. Chung, "Prediction model of user physical activity using data characteristics-based long short-term memory recurrent neural networks," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 4, pp. 2060–2077, 2019.

[5] H. Yoo and K. Chung, "Classification of multi-frame human motion using CNN-based skeleton extraction," *Intelligent Automation & Soft Computing*, vol. 34, no. 1, pp. 1–13, 2022.

[6] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris *et al.,* "Fall detection and activity recognition using human skeleton features," *IEEE Access*, vol. 9, pp. 33532–33542, 2021.

[7] C. Feichtenhofer, H. Fan, J. Malik and K. He, "SlowFast networks for video recognition," in *Proc. ICCV*, Seoul, SEL, Republic of Korea, pp. 6202–6211, 2019.

[8] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng *et al.,* "YOLOv6: A single-stage object detection framework for industrial applications," 2022. [Online]. Available: https://arxiv.org/abs/2209.02976

[9] S. Menshov, Y. Wang, A. Zhdanov, E. Varlamov and D. Zhdanov, "Simple online and realtime tracking people with new "soft-IoU" metric," in *Proc. AOPC*, Beijing, China, vol. 11342, pp. 148–154, 2019.

[10] P. Liu, M. Lyu, I. King and J. Xu, "SelFlow: Self-supervised learning of optical flow," in *Proc. CVPR*, Long Beach, CA, USA, pp. 4571–4580, 2019.

[11] H. Zhang, L. Xiao and G. Xu, "A novel tracking method based on improved FAST corner detection and pyramid LK optical flow," in *Proc. CCDC*, Hefei, Anhui, China, pp. 1871–1876, 2020.

[12] A. Aminfar, N. Davoodzadeh, G. Aguilar and M. Princevac, "Application of optical flow algorithms to laser speckle imaging," *Microvascular Research*, vol. 122, pp. 52–59, 2019.

[13] H. Yoo and K. Chung, "Deep learning-based evolutionary recommendation model for heterogeneous big data integration," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 9, pp. 3730–3744, 2020.

[14] H. Yoo, R. C. Park and K. Chung, "IoT-based health big-data process technologies: A survey," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 3, pp. 974–992, 2021.

[15] S. Park, J. Baek, S. Jo and K. Chung, "Motion monitoring using Mask R-CNN for articulation disease management," *Journal of the Korea Convergence Society*, vol. 10, no. 3, pp. 1–6, 2019.

[16] F. A. Khan, A. Gumaei, A. Derhab and A. Hussain, "A novel two-stage deep learning model for efficient network intrusion detection," *IEEE Access*, vol. 7, pp. 30373–30385, 2019.

[17] Z. Tian, C. Shen, H. Chen and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proc. ICCV*, Seoul, SEL, Republic of Korea, pp. 9627–9636, 2019.

[18] M. A. Haq, "CNN based automated weed detection system using UAV imagery," *Computer Systems Science and Engineering*, vol. 42, no. 2, pp. 837–849, 2022.

[19] A. Jawaharlalnehru, T. Sambandham, V. Sekar, D. Ravikumar, V. Loganathan *et al.,* "Target object detection from unmanned aerial vehicle (UAV) images based on improved YOLO algorithm," *Electronics*, vol. 11, no. 15, pp. 2343, 2022.

[20] M. A. Haq, M. A. R. Khan and M. Alshehri, "Insider threat detection based on NLP word embedding and machine learning," *Intelligent Automation & Soft Computing*, vol. 33, no. 1, pp. 619–635, 2022.

[21] S. Kumar, M. A. Haq, A. Jain, C. A. Jason, N. R. Moparthi *et al.,* "Multilayer neural network based speech emotion recognition for smart assistance," *Computers, Materials & Continua*, vol. 74, no. 1, pp. 1523–1540, 2023.

[22] G. Moon, J. Y. Chang and K. M. Lee, "Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image," in *Proc. ICCV*, Seoul, SEL, Republic of Korea, pp. 10133–10142, 2019.

[23] Z. Geng, K. Sun, B. Xiao, Z. Zhang and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *Proc. CVPR*, Nashville, TN, USA, pp. 14676–14686, 2021.

[24] X. Nie, J. Feng, J. Zhang and S. Yan, "Single-stage multi-person pose machines," in *Proc. ICCV*, Seoul, SEL, Republic of Korea, pp. 6951–6960, 2019.

[25] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. CVPR*, Nashville, TN, USA, pp. 16105–16114, 2021.

[26] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng *et al.,* "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[27] T. -Y. Lin, M. Michael, B. Serge, H. James, P. Pietro *et al.,* "Microsoft coco: Common objects in context," in *Computer Vision-ECCV 2014: 13th European Conf.*, Zurich, Switzerland, pp. 740–755, 2014.

[28] A. Doering, D. Chen, S. Zhang, B. Schiele and J. Gall, "PoseTrack21: A dataset for person search, multi-object tracking and multi-person pose tracking," in *Proc. CVPR*, New Orleans, LA, USA, pp. 20963–20972, 2022.

[29] X. Liao, K. Li, X. Zhu and K. R. Liu, "Robust detection of image operator chain with two-stream convolutional neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 955–968, 2020.

[30] W. Chen, Z. Jiang, H. Guo and X. Ni, "Fall detection based on key points of human-skeleton using openpose," *Symmetry*, vol. 12, no. 5, pp. 744, 2020.

[31] MetaAI, 2019. [Online]. Available: https://ai.facebook.com/research/publications/slowfast-networks-for-video-recognition/

[32] J. S. Chung, "Naver at ActivityNet challenge 2019—task B active speaker detection (AVA)," 2019. [Online]. Available: https://arxiv.org/abs/1906.10555

[33] V. Ramakrishna, D. Munoz, M. Hebert, J. Andrew Bagnell and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *Computer Vision-ECCV 2014: 13th European Conf.*, Zurich, Switzerland, pp. 33–47, 2014.

[34] S. E. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh, "Convolutional pose machines," in *Proc. CVPR*, Las Vegas, NV, USA, pp. 4724–4732, 2016.

[35] J. Arunnehru, S. Thalapathiraj, R. Dhanasekar, L. Vijayaraja, R. Kannadasan *et al.,* "Machine vision-based human action recognition using spatio-temporal motion features (STMF) with difference intensity distance group pattern (DIDGP)," *Electronics*, vol. 11, no. 15, pp. 2363, 2022.

[36] J. Yang and D. Claude, "An incipient fault diagnosis methodology using local Mahalanobis distance: Detection process based on empirical probability density estimation," *Signal Processing*, vol. 190, no. 4, pp. 108308, 2022.

[37] K. Will, C. Joao, S. Karen, Z. Brian, H. Chloe *et al.,* "The kinetics human action video dataset," 2017. [Online]. Available: https://arxiv.org/abs/1705.06950

[38] E. Cho, K. Sun and H. Yoo, "Real-time video analysis using SlowFast network based on short time-frame," in *Proc. KSII Transactions on Internet and Information Systems*, Jeongseon, Republic of Korea, pp. 163–164, 2021.

[39] AI Hub, 2023. [Online]. Available: https://aihub.or.kr/