IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 May 2023, accepted 30 May 2023, date of publication 2 June 2023, date of current version 9 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3282451

RESEARCH ARTICLE

Surveillance System for Real-Time High-Precision Recognition of Criminal Faces From Wild Videos

HYUN-BIN KIM[1], NAKHOON CHOI[1], HYE-JEONG KWON[1], AND HEEYOUL KIM[2]

[1]Department of Computer Science, Kyonggi University, Suwon 16227, South Korea
[2]Division of Computer Science and Engineering, Kyonggi University, Suwon 16227, South Korea

Corresponding author: Heeyoul Kim (heeyoul.kim@kyonggi.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2020R1A6A1A03040583.

ABSTRACT As violent criminals, such as child sex offenders, tend to have high recidivism rates in modern society, there is a need to prevent such offenders from approaching socially disadvantaged and crime-prone areas, such as schools or childcare centers. Accordingly, national governments and related institutions have installed surveillance cameras and provided additional personnel to manage and monitor them via video surveillance equipment. However, naked-eye monitoring by guards and manual image processing cannot properly evaluate the video captured by surveillance cameras. To address the various problems of conventional systems that simply store and retrieve image data, a system is needed that can actively classify captured images in real-time, in addition to assisting surveillance personnel. Therefore, this paper proposes a video surveillance system based on a composable deep face recognition method. The proposed system detects the faces of criminals in real time from videos captured by a surveillance camera and notifies relevant institutions of the appearance of criminals. For real-time face detection, a down-sampled image forked from the original is used to localize unspecified faces. To improve accuracy and confidence in the recognition task, a scoring method based on face tracking is proposed. The final score combines the recognition confidence and the standard score to determine the embedding distance from the criminal face embedding data. The blind spots of surveillance personnel can be effectively addressed through early detection of criminals approaching crime-prone areas. The contributions of the paper are as follows. The proposed system can process images from surveillance cameras in real-time by using down-sampling. It can effectively identify the identity of criminals by using a face tracking ID unit and minimizes prediction reversal by solving the congested embedding problem in the feature space that may occur when performing identification matching on a large amount of face embedding DBs. Additionally, the reliability of the identification results is complemented by an identification score accumulation method. In this paper, we prototyped the proposed system and experimented with the recognition model, achieving an accuracy of 0.900 and an F-1 score of 0.943. We also experimentally confirmed that the models proposed in other studies have higher performance when using the tracked instance-level face identification method proposed in this paper. It is expected that the proposed system can be used to locate criminals and protect national facilities, and such responses can quickly prevent accidents/incidents. The dataset and code are available at *https://github.com/aengoo/focusface*

INDEX TERMS Crime prevention, down-sampling, face recognition, video.

I. INTRODUCTION

Video surveillance systems have long been used and evolved from analog to digital systems. With the heightened social interest in crime prevention and public safety, image security and surveillance equipment systems have sharply increased, and high-performance, low-cost imaging devices have been introduced [1]. However, such systems focus on an administrator conducting naked-eye monitoring of videos captured by the surveillance cameras. Moreover, providing effective

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Sanchez.

operations is also difficult because the control systems and the personnel for monitoring and managing the collected videos are limited. Images or videos captured by a surveillance camera are stored for 3–15 days depending on the environment of the system and are renewed in a first-in, first-out order. Most videos do not serve a meaningful purpose and are deleted in the process of storing new videos. The images captured by security devices do not have meaningful functional utilization unless used for crime prevention and public safety, such as in the appearance-based identification of ex-convicts of specific crimes with high recidivism rates. In particular, as sexual crimes targeting the socially disadvantaged (including minors) tend to have a high recidivism rate [2], ex-convicts must be strictly monitored and restricted from approaching crime-vulnerable areas, including schools and childcare centers.

The Korean government uses electronic anklets [3] to track the locations of former convicts. However, there must be another method for tracking ex-convicts approaching crime-vulnerable areas when such tracking devices are arbitrarily removed. Accordingly, the installation of surveillance cameras in crime-vulnerable areas has become mandatory following childcare protection, which has recently been emphasized. Thus, the national government and relevant institutions are securing resources and establishing dedicated operation bodies for managing video data through an integrated center for controlling surveillance cameras [4]. Despite such efforts, most of the infrastructure requires simple eye monitoring by available personnel. The crime-vulnerable groups and regions can be effectively protected from criminals if the videos captured by the installed surveillance cameras are searched in real time to actively detect criminals and notify the relevant institutions.

This study proposes an image security device control system based on deep-learning facial recognition. To overcome the limitations of naked-eye monitoring in conventional systems, the proposed system aims to automatically and accurately recognize criminals appearing in crime-vulnerable areas in real-time and to notify relevant institutions, thereby preventing crimes. A deep neural network-based feature extraction method and feature distance comparison method in the feature space are used to verify criminals appearing in videos captured by a fixed imaging device, and the identification information of criminals is examined through a series of identification processes. Unlike facial recognition within a fixed range of access control, the proposed system aims to detect and identify criminal faces accompanied by occlusions or avoidance reactions in videos containing a large number of moving people. The proposed system requires high processing performance for real-time applications and is achieved through a down-sampling technique.

The facial recognition in the proposed system is conducted in two steps: face detection and face identification. In the conventional identification processes, predictions are generally made through frames using specific distance functions, but processes involving distance functions cannot disregard outliers that may intermittently occur, leading to incorrect prediction results. In this study, the prediction results regarding the appearances of criminals are derived at each frame, and the identification scores of each face are accumulated through object tracking to minimize the overturning of previous predictions. By adjusting the weights of the accumulated identification scores, a standard score and detection confidence for the identification are used to exclude abnormal face identification results. This study proposes a scoring method based on face tracking such that the final score combines the detection confidence and standard score for the embedding distance, as determined from criminal face embedding data. This improves the accuracy of the identified faces and increases confidence in face recognition in the identification step.

The deep learning-based facial detection process is highly accurate but slow and takes up most of the computation time of the overall system. Both the fixed and accumulated latency must be considered when performing face detection from video data in real-time. It is impossible to eliminate the fixed latency. Nevertheless, the accumulated latency must be avoided, because the generated data can exceed the processing speed of the system. The frames per second (FPS) metric represents the processing speed. It is a crucial indicator of the computation speed of a system and is related to the accumulated latency. Improving the FPS is an important task for ensuring that the processing delays of input videos do not accumulate. Several studies have been conducted on improving FPS image batch processing, thereby taking advantage of advances in parallel processing. To perform batch processing for an image stream in real-time, however, there is a delay until a certain number of frames are captured, ultimately leading to an increased fixed latency time. The system proposed in this study consists of multiple processes as a procedure, and additional delays occur in the object tracking step because the data are sequentially checked. Therefore, applying batch processing methods for videos was deemed inappropriate for the proposed system. However, the accumulated latency can be prevented by ensuring that the individual frames are processed without delays when batch processing is not applied.

The proposed system improves the processing speed by utilizing down-sampling only for the images in the face detection process. It is difficult to extract identifiable features from a face image with insufficient resolution. Therefore, in the identification task, a face image extracted from an image with as high of a resolution as possible is required. However, the face localization task requires a relatively low-resolution image. It is possible to apply down-sampling only in the detection process because it is possible to filter out prediction errors through object tracking for checking whether the detection results in successive frames are continuous. The face detection process occupies more than half of the total processing time of this system, so even considering the time required for the down-sampling, the efficiency gain from using a lower resolution is very large. By using temporarily down-sampled frames during the detection process, the processing speed reaches over 30 FPS, far exceeding the 15 FPS refresh rate of general surveillance cameras.

The proposed system aims to overcome the limitations of the existing surveillance system, which relies on human monitoring capabilities. The proposed system ensures that video footage from security cameras is thoroughly inspected, allowing for the early identification of potential criminals or threats in crime-prone areas such as schools, daycare centers, and national security facilities. This contributes to enhanced safety by enabling prompt notifications to relevant authorities through the surveillance system's notification pipeline. In Chapter 4, we demonstrate a real-time crime information system that analyzes surveillance videos to detect, alert, and prevent potential incidents such as missing children, criminal activities, and threats to national facilities.

This paper proposes various new techniques for implementing a recommendation system and presents the results of experiments. The main contributions of this paper, according to these methods, can be summarized into four points as follows:

1) A face detection method is proposed for processing video frames in real-time through down-sampling.
2) The potential problems concerning the prediction unit when performing facial recognition on video data are outlined, and an identification method with a face tracking ID unit is suggested for effectively recognizing criminals from video data.
3) Prediction overturns are minimized by addressing the congested embeddings arising in a feature space that occur while performing identification on a large-scale face embedding database (DB). In addition, an identification score accumulation method for outputting only highly reliable identification results is proposed.
4) An evaluation dataset is created to verify the system operation performance under a fixed high-resolution video input condition. The proposed system demonstrates faster and more accurate facial recognition performance for the same dataset relative to existing systems in which processes are interrelated through simple connections.

The remainder of this paper is organized as follows. In Section II, the related work on crime-prevention systems is introduced, along with concepts regarding general facial recognition from videos. In Section III, the proposed system and model are explained in detail. In Section IV, the experimental environment, dataset, and implemented system demonstration are described. Finally, conclusions are drawn in Section V.

II. RELATED WORKS

A. SURVEILLANCE-BASED CRIME PREVENTION SYSTEM

1) CASES OF CRIME PREVENTION USING SURVEILLANCE CAMERA

The presence of surveillance cameras in public areas can suppress crime. Surveillance cameras, which are considered for situational crime prevention, indicate that the official monitoring level of a target region is heightened [5]. Piza et al. [6], [7] reported that the installation of surveillance cameras in residential areas provides an effective measure for crime prevention. Alexandrie et al. [8] also confirmed that the crime rate decreased by 24–28% in public places based on seven types of natural experiments involving surveillance cameras. A crime is defined as an "intentional behavior for gratifying a criminal's everyday needs," [9] and criminals consider several aspects while executing specific crimes, including the inherent risks and compensation for potential need gratification. The installation of surveillance cameras adds inherent risks to a criminal's selected behavior and urges observers to be more conscious, thereby triggering a perception mechanism in persuading criminals not to commit a crime [10].

Research is consistently being conducted on increasing the utilization of the images or videos captured by surveillance cameras. Some studies have examined methods for restoring images captured in inclement weather or that are difficult to analyze owing to shading, whereas other studies have focused on the structures in which different types of sensors are integrated (infrared camera, pan-tilt-zoom camera, vision camera, etc.).

Recently, research has been conducted on integrated crime prevention systems based on big data and the Internet of Things [11]. The trend is evolving from crime prevention toward environmental design, which ranges from physical environment design to ubiquitous crime prevention including sending warnings using non-structured social network data. However, conventional systems cannot actively interpret the information from linked information collectors (such as image surveillance devices), and thus criminal situations are only identified after crimes have already happened or after the risk of an imminent crime has been confronted.

2) CRIME PREVENTION USING COMPUTER VISION

All countries worldwide currently attempt to hinder the recurrence of crimes through legal punishments, aiming to ultimately prevent the occurrence of crimes in the first place. The types, methods, and frequency of criminal activities are soaring as the population is sharply increasing and a larger number of individuals can easily obtain information; hence, the demand for a predictive system is on the rise [12]. During a criminal investigation, the analysis is generally performed using visual data. For crime prevention activities, computer vision can provide significant help by providing automated extraction of the most essential data from the unnecessarily vast amount of available data. Major tasks using computer vision related to crime prevention and investigation include license plate recognition [13], action and posture recognition [14], and facial recognition [15]. These tasks typically require analyzing 2D images, but a growing number of studies have been conducted using visual data of three dimensions or more [16]. Among the 3D data processing techniques related to facial recognition, 3D face restoration reconfigures a face into a 3D shape model or mesh from the given data [17]. Video data can also be utilized by considering a dimension expansion from the temporal perspective rather than the spatial perspective. Videos typically undergo excessive lossy compression owing to the large storage space necessary for storing the videos. Moreover, surveillance footage frequently does not help investigations owing to the poor quality of data. The research was recently conducted to improve the resolution of surveillance footage based on a generative adversarial network (GAN) [18]. A super-resolution GAN (SRGAN)-based technique improved the resolution of the main objects in images, but the resolution of faces, which often are extremely small objects with delicate features, could not be restored. This is based on the grid-level subsampling used in most image compression algorithms. In these, the detailed representations of objects smaller than the grid become severely damaged; thus, restoring the resolution after capturing images with super-resolution is considered as difficult. As mentioned above, research on crime prevention through computer vision is being conducted in various fields. We propose a crime prevention solution through face recognition among these vision-related research fields. Our research for crime prevention uses techniques for the detection and recognition of images and image quality adjustment.

B. FACIAL RECOGNITION

1) GENERAL FACIAL RECOGNITION

In general, a facial recognition system applies a three-step process of face detection, face feature extraction, and face identification or verification. This approach demonstrates a similar pipeline to the 2-stage object detector. In the case of face recognition problems, using the 2-stage method is inevitable because training a FCN-based classifier simultaneously, as in the case of a 1-stage detector, is unsuitable for actual use. On the other hand, PCA-based methods can quickly modify classifiers without the need to retrain the entire network. Therefore, recent studies have constructed 2-stage recognizers for PCA-based face identification or verification methods on Pretrained CNN Features, following CNN-based face detection [19]. In the face detection step, a face of a person is initially detected from the entire input image. Then, face feature extraction is performed to compare the inherent features of each face. There have been significant performance improvements for a large number of computer vision problems, owing to deep learning-based machine learning techniques and advances in hardware. Most research on facial recognition systems uses deep learning. Xiang et al. [20] achieved a face identification accuracy of 99.10% on the MegaFace dataset, whereas Zhu et al. [21] achieved a face detection average precision (AP) of 0.924 on the hard subset of the WIDER FACE dataset. However, in these studies, the accuracy level was based on experimental observations of the respective dataset; therefore, there is a possibility that the recognition accuracy could be lower for individuals who demonstrate recognition avoidance, such as criminals in real life. In this section, the research trends in each sub-process of the latest facial recognition systems are explained in further detail.

Facial recognition begins by extracting the face area and features, followed by property extraction or occlusion correction. The face area extraction corresponds to a classification problem for determining whether the detected part of the image is a face. The Viola-Jones object detection framework [22] using the boosted Haar cascade classifier demonstrates stable performance; the histogram of oriented gradients (HoG) is also widely used. Regarding deep learning-based technologies, several methods, including the multi-task cascaded convolutional neural network (MTCNN), MobileNet-SSD, and You Only Look Once-Face (YOLO-Face), have been proposed to solve the tiny-object problem and to ensure stable performances in face detection under diverse environments and conditions.

The faces extracted through face extraction enable the differentiation of each individual. Facial recognition aims to find faces within a certain range based on a threshold value, or by comparing the mutual similarities with elements within a facial search space created in advance. Different methods are used for the similarity comparison, based on the extent of the computations. For higher accuracy, techniques such as the principal component analysis (PCA) and independent component analysis (ICA) [23], [24] are often used in magnetization as a function of the applied field. The amount of computation can be decreased by reducing the dimensions of the images. Among deep learning-based methods, artificial neural networks such as autoencoders embed facial information into a lower-dimension vector to reduce the dimensions. Then, the similarity between faces is determined based on computations with embedded vectors.

Ben-Baruch et al. [25] trained MobileNet using Knowledge Distillation to meet the demand for lightweight models. Alansari et al. [26] presented a face feature extraction network based on GhostNet, which linearly replicates duplicated features, as another study on lightweight models.

2) FACIAL RECOGNITION IN VIDEOS

The facial recognition of random individuals appearing in videos may pose ethical and practical issues. To avoid such issues, the system must be capable of recognizing unspecified individuals without mistaking innocent pedestrians for criminals and while demonstrating a high recall rate. In the video data captured by surveillance cameras, frames are continuously captured within a certain time interval. General smartphones or video cameras can capture at rates of 30 or 60 FPS, whereas certain models can support up to 240 FPS. Existing surveillance cameras serve the purpose of image recording, and typically save at a rate of 15 FPS owing to their limited storage space; however, actual cameras can capture at higher speeds. Videos often represent higher-dimensional data than static images and display fragmentary information. Because static images can be very easily distorted, the latest methods utilize information in higher dimensions to obtain highly reliable results. Artificial neural networks can be extended to consider the possibility of distortion in 2D data into 3D data or 3D restoration through geometric estimations, such as in stereo vision [27]. Recently, depth maps with arbitrary dimensions and 3D data through a point cloud were secured using infrared reflected light (such as LiDAR [28]. As the range of methods capable of addressing real-time problems remains very limited, methods requiring a large amount of time to analyze each frame are difficult to adopt. A system operating in real-time requires a minimum computation time. If face tracking is used in videos, high-dimensional data can be quickly secured for face identification.

The mathematical background of FaceNet, which is the basis of our proposed model, mainly focuses on the triplet loss function and optimizing the distance between faces in a high-dimensional embedding space. This allows us to quantify the similarity between face images to perform recognition and verification tasks. Triplet loss uses three face images: a reference image (Anchor), an image of the same person as the reference image (Positive), and an image of a different person than the reference image (Negative). The embedding function is called f(x), and the embedding for each image can be denoted by $f(A)$, $f(P)$, and $f(N)$. The goal of triplet loss is to reduce the distance between pairs of positives and increase the distance between pairs of negatives. To accomplish this, the conditions in Equation (1) must be satisfied. α in Equation (1) is the margin, which represents the minimum difference between the positive and negative distances.

$$\|f(A) - f(P)\|^2 + \alpha \leq \|f(A) - f(N)\|^2 \tag{1}$$

$$L(A, P, N) = max\left(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0\right) \tag{2}$$

The triplet loss function is defined as shown in equation (2). During model training, we update the weights to minimize this loss function. As the optimization progresses, the distance between faces in the embedding space is optimized, bringing faces of the same person closer together and faces of different people farther apart. In conclusion, the mathematical background of FaceNet is based on the triplet loss function, which measures the similarity between face images by optimizing the distance between faces in the embedding space, and provides the basis for recognition and verification tasks. We contribute to achieving high accuracy in face recognition systems by advancing the image frame-level FaceNet to video-level recognition.

Human visual perception uses significantly more information than what can be captured by a typical camera. For example, two eyes can be used to perform stereo-vision based depth estimation. Additionally, longer viewing times while continuously gazing at a subject can lead to more accurate visual judgments. In our research, we focused on the latter case of gaze fixation. However, we were unable to find any prior research on this idea.

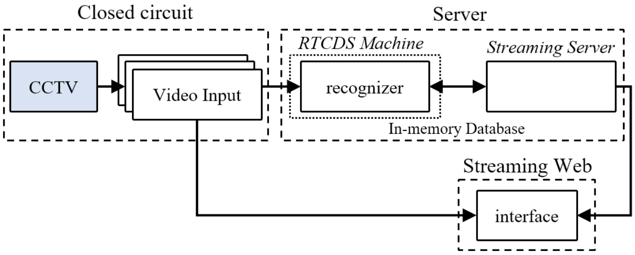


FIGURE 1. Overall system architecture.

We use PCA-based deterministic predictive models for face identification in a frame-based face recognition pipeline. In continuous video frames, we track related face instances according to a Markov decision process for prediction. This allows for more precise predictions based on cumulative probability models.

III. PROPOSED METHODS

A. SYSTEM OVERVIEW

This study proposes a real-time criminal detection system (RTCDS) for high-risk ex-convicts based on face tracking utilizing a deep learning-based face detector and identifier. The proposed system aims to prevent crimes by detecting a face in video data captured by image security devices such as surveillance cameras, and then notifying a chief manager or relevant criminal institutions after identifying the criminal through a comparison with criminal DBs. The overall architecture is shown in Figure 1.

Figure 2 illustrates the process structure of the proposed face detection and recognition system. The system consists of general processes marked with initials in Figure 2, as well as unique processes proposed in this study. Similar to a general facial recognition system, the proposed system consists of face detector D, tracker T, face encoder E, and identifier I [29]. The proposed system must perform the analysis with a minimum latency time by having each frame of the captured video data immediately pass through the processing sequences.

The following two unique processes are employed.

1) Sampling: The processing latency is reduced by using low-resolution samples during the face detection, whereas high identification accuracy is maintained during the identification step by cropping the face regions from the original high-resolution images.
2) Score Dictionary Identification: Scores related to the identification results are accumulated for the face ID of each tracked face by creating a dictionary with the tracking ID as a key. The dictionary monitors the score such that when the score exceeds an arbitrary threshold, the system notifies parties of the identification result and transmits the appearance of the criminal.

The face detector D predicts the position and size of each face to be identified from each frame image. A binary detection (or localization) framework that outputs the confidence

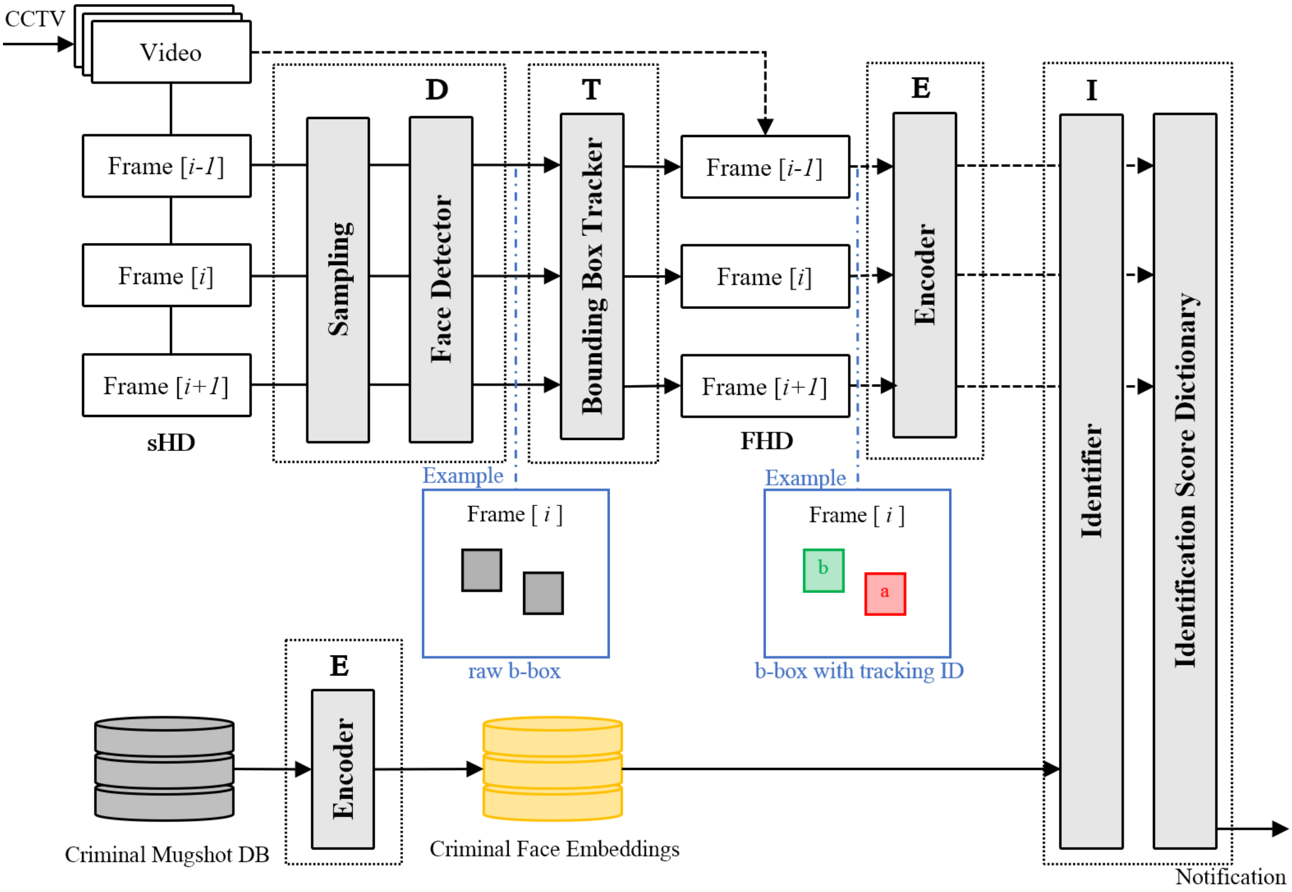


FIGURE 2. Structure of the proposed system for face detection and recognition.

score and bounding box must be used as the detector; this condition is satisfied by most general object detection frameworks. The most critical point in selecting a face detector D for the proposed system is the size of the receptive field. Most feature-extraction networks have fixed receptive fields. Most high-performance feature extractors with an appropriate level of latency for real-time processing only receive images with a resolution of 224 × 224. Certain models receive images with higher resolutions, such as 384 pixels; however, as the resolution increases, the depth of the feature extractor and latency also increases. Because a face detection framework using a deformable convolutional network is free from such input resolution limitations, a sufficient amount of information can be obtained from the high-resolution images with relatively lower latency compared to other networks with depths proportional to the input resolution. Therefore, ''RetinaFace'' [30] was selected as the most appropriate face detector for the proposed system.

The face tracker T assigns a tracking ID to each bounding box to indicate the same person in each frame before identifying the detected faces. This process is required for combining the confidence score and the class identified in different continuous frames for each tracking ID. Deep learning-based detection and identification must be performed within 30 ms for real-time processing, implying that the face tracking must be processed very quickly concerning the time required for inputting and outputting images. A simple online real-time tracker (SORT) [31] satisfying these conditions and providing stable performance was selected as the instance tracking method. Using deep SORT [32], an advanced extension of the SORT was also considered. This method can effectively solve realistic problems arising from the occlusions occurring in the SORT. In contrast, a face detector is trained for binary detection and thus cannot extract effective feature information for an intra-class classification between face instances. Thus, there is an insignificant advantage to using deep SORT in the proposed system instead of SORT in terms of performance while sacrificing computation time.

The face image encoder E and identifier I are based on the method proposed by Schroff et al. [33]. FaceNet is the baseline model; it is a metric learning-based technique in the facial recognition field. Therefore, the basic model of the proposed system demonstrates a significantly improved identification accuracy by applying the FaceNet-based encoder and identifier compared to using the basic encoder-identifier.

Figure 3 intuitively shows the overall operation flow of the proposed system. It starts with frames obtained from surveillance camera devices such as security CCTV. Face

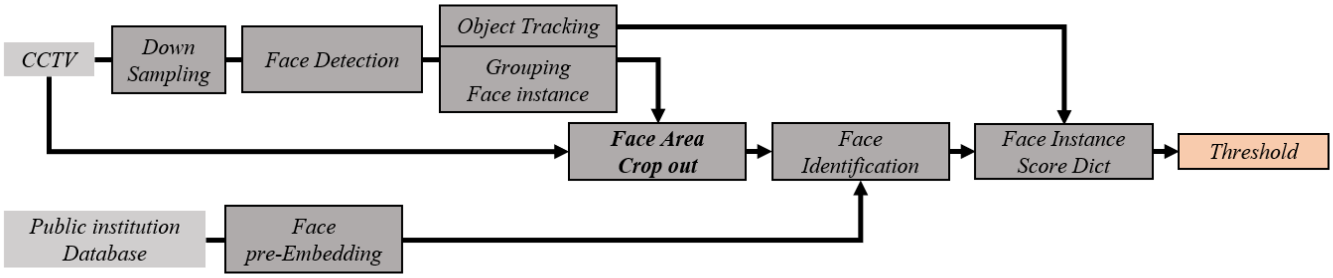


FIGURE 3. Overall process pipeline of the proposed method.

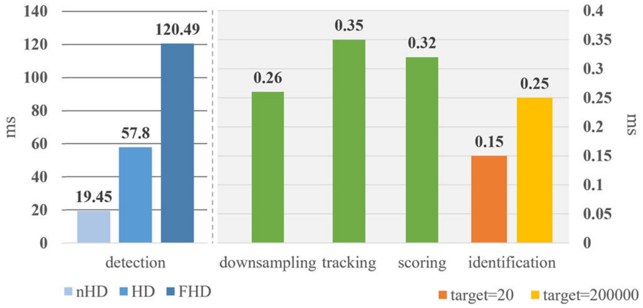


FIGURE 4. Comparison of processing latency for each process in the proposed pipeline.

detection is performed from the frame after down-sampling. Then tracking of face instance is performed and the identification score is calculated. This score is accumulated, and the system reports the recognition result if the score exceeds the threshold. This processing pipeline iterates continuously. Figure 3 abstracts the system structure shown in Figure 2 in the form of a process pipeline.

B. REAL-TIME TWO-STAGE RECOGNITION

Most conventional artificial neural network models (including CNNs) have a limited receptive field for each model. Accordingly, images with unsuitable resolutions must be resized in advance. This is mostly an issue in terms of implementation, as the process is determined according to the size of the receptive field of the network being used. In contrast, the face detection network applied in the proposed system is applied with a deformable convolutional network. As such, it places no restrictions on the resolution of the input images. Therefore, the most appropriate resolution for the detector must be determined from the perspective of the usability of the overall system while recognizing the trade-off between the detection performance and speed. This section explains the sampling process shown in Figure 2 in detail.

The RTCDS system aims to recognize the faces of persons appearing in unrefined video data. The position of each instance needs to be localized for the same reason as that for the tasks used in instance segmentation or general object detection. Face detection is often performed using a traditional handcrafted feature extraction method or convolutional networks. Traditional techniques capable of quick computation are commonly used in this context, owing to the real-time processing constraint [34]. However, the detection accuracy of most systems using traditional detection techniques are unreliable.

Furthermore, the detector must satisfy performance criteria at a certain minimum level, as the operating performance of an object tracker is heavily dependent on the performance of the detector. RetinaFace is a face instance detection framework in which the scale-invariant performance is assured by applying a feature pyramid network [35] and deformable convolutional network (DCN) [36] to construct a general deep CNN as the feature extractor.

The proposed method outperforms other similar techniques based on a multi-task training strategy, and it has a short processing latency as a single-stage detector. RetinaFace has no restrictions on the resolution for the input about the structural characteristics of the framework, but the processing time increases in proportion to the area of the input images. Considering the characteristics of the DCN, there is no network processing latency in terms of the resolution. This is because the resolution and network depth do not correlate unlike in general feature extraction networks; nevertheless, they contribute to the post-processing time, which is affected by the resolution. The detection accuracy decreases as the resolution is reduced, but the trade-off in regard to the resolution reduction is not significant above a certain level. An experiment was conducted to verify the minimal feature information required to confirm that a face could be included in images with 380-pixel resolution. The latency was drastically reduced according to the area reduction ratio when images with full-high-definition (FHD) resolution were down-sampled through bilinear interpolation to ninth high-definition resolution, i.e., 640 × 360 as can be seen from the leftmost graph in Figure 4.

Figure 4 also shows the estimated latency for each process in the proposed pipeline model. Most of the latency is caused by the detector, whereas the other processes only require less than 1 ms of latency. This characteristic of our pipeline highlights the advantages of using temporal down-sampling. Temporal down-sampling in the detection process drastically reduces the latency and has a critical positive impact on the overall pipeline performance. The bounding box precision

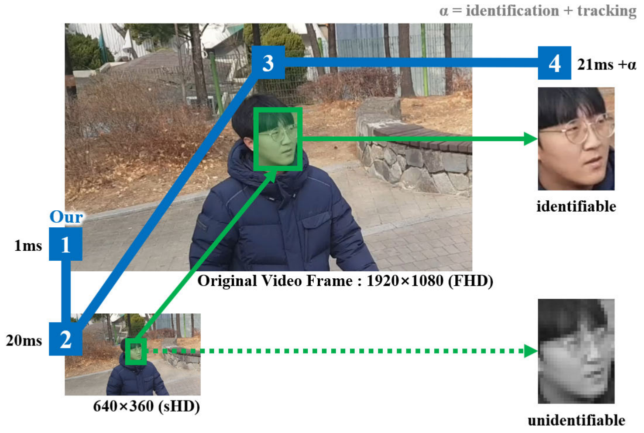

α = identification + tracking
3
4
21ms +α
identifiable
Our
1ms
1
Original Video Frame : 1920×1080 (FHD)
20ms
2
640×360 (sHD)
unidentifiable

FIGURE 5. Flowchart of a proposed system for using appropriate resolutions for each process.

degradation owing to the down-sampling is improved by the subsequent object-tracking process.

It is also worth noting that the proposed model handles input video in real-time because the overall latency per frame is less than 33ms as can be seen in Figure 4. Of course, there is an apprehension that the latency may increase when increasing the number of categories to be identified, that is, the number of target persons to be identified. However, in the identification process, the number of targets has little impact on the latency. As the Euclidean distance calculation between 128-dimensional vectors can be vectorized, the actual operation speed hardly slows down even if hundreds of thousands of targets are set, as can be seen from the rightmost graph in Figure 4. The link below also provides a supplementary demo video showing the real-time processing capability of the proposed model.

Link: https://github.com/aengoo/focusface

Figure 5 explains the method for applying down-sampling to the video frame inputs in real-time as proposed in our system. The proposed method proceeds in the following order: (1) Images captured in full high-definition (FHD) resolution are down-sampled to the ninth high-definition (nHD) resolution. (2) The face position is searched in a low-resolution image, which takes an average of 20 ms. (3) A high-resolution face image is cropped by projecting the detected face area onto a high-resolution image. (4) Identification is performed by encoding the face image. In this process, the time required for identification and tracking is set to α, and the total processing time excluding α is measured to be 21ms on average. The detection accuracy improves when a high-resolution image is input into the face detection network. However, the detection accuracy increases with the resolution improvement as a log function, in which the improvement range becomes extremely small beyond a certain resolution threshold. Therefore, the face position was detected by down-sampling the image to the minimum resolution that would allow the face position to be verified, and the detected region was extracted from a high-resolution image in the subsequent step to ensure highly accurate identification.

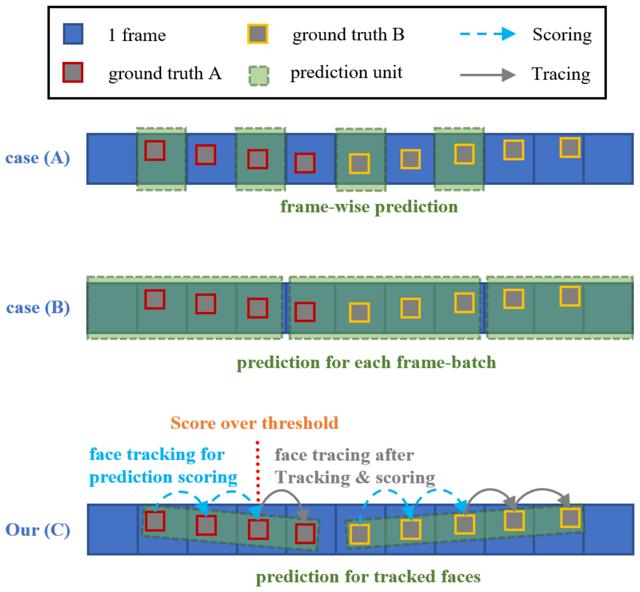

1 frame
ground truth B
Scoring
ground truth A
prediction unit
Tracing
case (A)
frame-wise prediction
case (B)
prediction for each frame-batch
Score over threshold
face tracking for prediction scoring
face tracing after Tracking & scoring
Our (C)
prediction for tracked faces

FIGURE 6. Examples of three different methods for facial recognition in videos: (A) Individual frame prediction, (B) Prediction of frame batch captured at a certain time interval, (C) Prediction of detected and tracked face object.

C. APPEARANCE RECOGNITION SYSTEM

In general, video data consists of multiple static images captured within a certain time interval. In other words, video data comprises a set of images sampled at discrete times. The following prediction procedure may be followed to recognize the face of a person in the video data. First, based on the output image recognition results for each frame, there is a high possibility for the prediction results to be overturned for the same face instance unless the identification accuracy of the model is extremely high. The problem of prediction being overturned is more likely when the prediction accuracy is lower owing to the image quality, low lighting, or imaging distance. Second, a prediction method for frame batch is used. A recurrent neural network architecture such as a long short-term memory or 3D CNN can be used for this method. This method encounters the problem of determining the size of a unit set to allow inferences to be obtained in situations where images are continuously input in real-time. Furthermore, the confidence level differs depending on whether the inference is being made for a case with the face lying on the border of the prediction unit versus at the center of the prediction unit.

Accordingly, the following two conditions are imposed on our system.

1) The prediction results should not be overturned.
2) The prediction confidence for a face instance should not be affected by where the face appears.

To satisfy both of the aforementioned conditions, a method is proposed in which the data being analyzed are grouped according to the tracked face object ID, and then prediction results are output at the group level. Object tracking in general is a technique where detection continues even when the region of interest is divided and deformation, occlusion, and position changes occur; various techniques including frame differencing, optical flow, and background subtraction can be used. These methods entail a unique object detection process for each tracking technique. The goal of object tracking is to assign the same ID to the face bounding boxes of the same person. In this context, the previously explained methods are not applicable, as they do not handle the bounding box detection format. The procedure starts with the face detection process, and then the intersection-over-union (IoU) distance between face regions detected in nearby frames is calculated to assign the same ID to a pair of face regions within a certain IoU distance. If this process is repeated for each frame, the tracked objects can be grouped according to their IDs.

Figure 6 illustrates three prediction approaches for a given video frame. In each prediction, the red and yellow rectangles represent the prediction targets, and the green border is the unit of each prediction. The blue arrows represent the cumulative sequence of predictions until the score is exceeded, and the gray arrow indicates face tracking after the prediction result is output. As shown in Figure 6, the facial regions are identified in each frame, and the identification results for the tracked facial objects are accumulated by their IDs. When the accumulated prediction score exceeds a threshold, the final prediction result is output as shown in case (C). This method has the advantage that the prediction results for facial object IDs are never reversed. To output the correct prediction results, we also provide a method for calculating the confidence level of the accumulated prediction results. This will be explained in detail in Section III-D.

The object detector based on the bounding box applied to the system is based on the SORT. Because the time interval between frames in normal 30 FPS images is approximately 33.33 ms, the probability P of overlap between the bounding boxes of a face in two consecutive frames of videos captured at a fixed camera position is very high. The detected face regions must be expressed as bounding boxes. If the IoU distance between the face bounding boxes in two consecutive frames is calculated, the probability P denotes the probability that two bounding boxes with the minimum IoU distance show the faces of the same person. The computation time can be considerably reduced if the Hungarian algorithm is applied because the process of finding overlapping bounding boxes is 1:1 when matching the elements of the two sets. In addition, tracking can be continued even if faces cannot be detected in certain frames (e.g., because of occlusion or deformation) by calculating the moving acceleration of the bounding boxes based on a Kalman filter.

D. WEIGHTING THE PREDICTION CONFIDENCE

When the tracked face objects are predicted as a group unit as explained in Section III-C (3.C), the results may not be output if the face never disappears in the videos. Therefore, the appropriate timing for outputting the accumulated identification results must be determined to solve such problems. In this section, the operational scheme of the process corresponding to the score dictionary in Figure 2 is explained.

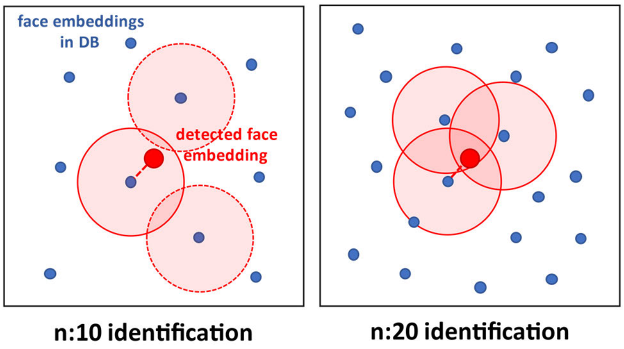


FIGURE 7. Illustration of how vector distances become denser as the number of faces for comparison increases in a restricted vector space. Blue dots show encoded facial embedding vector samples in the database, while red dots show detected facial embedding samples during system execution.

This system employs the method proposed in FaceNet as the baseline technique for face identification, as follows. First, embedding vectors with the feature information of the face images are extracted through the encoder. The Euclidean distance between the face embedding extracted from the videos and that in the criminal face DB is calculated to deduce identification when the distance less than the threshold is the same as the distance of the closest embedding. This distance metric-based identification technique is simple and intuitive. However, the distance metric cannot be the only source of confidence when the system is being applied to real tasks, because other issues may arise depending on the number of criminals in the DB. The distance between embeddings becomes closer in the embedding feature space as the number of criminals increases. The principle of the distance metric-based embedding identification can be explained as shown in Figure 7, where a face is detected when an embedding is input and positioned within the identification induction range of a circle having the distance threshold as a radius. As the distance threshold is fixed, a decreasing distance between embeddings indicates that the intersection of the identification induction range increases; the intersection area ratio also increases. Here, if the absolute face distance values are the same when the face embedding is within one embedding identification induction range and when located at the intersection of multiple embedding identification induction ranges, the confidence level varies. Therefore, the z-score of the distance score shown in the equation is used to discern reliable identification results without modifying the threshold. That is, the z-score is used to discriminate between ambiguous face embeddings. In addition, the more the face is head-on-photographed, and the clearer the face, the higher the reliability of the identification result. The confidence score for each bounding box resulting from the face detection process tends to have a higher value, as the face part is clearer and the face is photographed from the front. Therefore, we retain the confidence score from the detection process up to the identification process and add it to the final identification result. Finally, we calculate a reliable frame-by-frame identification score by joining the three indicators as shown in Equation (11). The detection network and post-processing operations are commonly referred to as $f_{det}\left(x\right)$ and $b^*_{i,j}$, respectively. $f_{det}\left(x\right)$ outputs the coordinates of the bounding box detected at x^i of the down-sampled i-th frame and detection confidence t_{conf}. The coordinates are converted to accommodate the resolution of the image used in the identification process. The respective areas are cropped, as each coordinate is predicted with a normalized coordinate. As shown in Equation (6), $f_e\left(x_{i,j}\right)$ expresses a d-dimension feature vector extracted from a cropped image patch at the original resolution at the position of each bounding box j by the encoder f_e. The down-sampled input image and related results are marked with the $*$ symbol. In this section, the $\bigcirc$-series operators $\ominus$, $\odot$, and $\oslash$ signify operations between corresponding elements, such as the Hadamard product $\odot$.

$$b^*_{i,j} := \left\{t^*_x, t^*_y, t^*_w, t^*_h, t_{conf}\right\}_{i,j} \in f_{det}\left(x^*_i\right) \tag{3}$$

$$b_{i,j} = \begin{pmatrix} t^*_x & t^*_y \\ t^*_w & t^*_h \end{pmatrix}_{i,j} \odot \begin{pmatrix} s_w & s_h \\ s_h & s_h \end{pmatrix} \tag{4}$$

$$x_{i,j} = \left(x_i\right)_{t_x:t_x+t_w, t_y:t_y+t_w} \tag{5}$$

$$f_e\left(x_{i,j}\right) \in R^d \tag{6}$$

In the above, $f_e\left(x\right)$ is a function for encoding specific face images. The face image $x_{i,j}$ is received from the original high-resolution image as input. The encoder $f_e\left(x\right)$ is expressed as a d-dimensional set of real numbers. It is assumed that the system encodes the face images of a criminal list before the operation, and the set Y in Equation (7) represents this encoded face image list.

The Euclidean distance $D\left(x_{i,j}\right)$ from all criminal face vectors is calculated and then the face similarity score $S\left(x_{i,j}\right)$ is calculated from $D\left(x_{i,j}\right)$ with a maximum embedding distance criteria τ; closer distances are converted to higher scores as shown in Equation (8). The criteria τ is set to 0.6 which is borrowed from [37]. Then, the standard score (z-score) set $S_z\left(x_{i,j}\right)$ is calculated using $S\left(x_{i,j}\right)$.

$$D\left(x_{i,j}\right) = \left\{\left|f_e\left(y_k\right) - f_e\left(x_{i,j}\right)\right|_2^2 |\forall y_k \in Y\right\} \tag{7}$$

$$S\left(x_{i,j}\right) = \tau \ominus D\left(x_{i,j}\right) \tag{8}$$

$$S_Z\left(x_{i,j}\right) = \frac{\left(S\left(x_{i,j}\right) \ominus \mu S\left(x_{i,j}\right)\right)}{\sigma S\left(x_{i,j}\right)} \tag{9}$$

The z-score of the face embeddings represents the relative similarity of each embedding, whereas the face distance represents the absolute embedding distance. When the z-score is used alone, faces captured as too small or sideways fail to display the minimum amount of information necessary for identification. The confidence score is output through the SoftMax function during the face detection process, and then the detection score and z-score are combined in consideration of the influences of the sizes and directions of the captured images on the confidence score. Furthermore, the embedding distance is also applied as an important metric for identification, and therefore, it is converted to a distance score to be summed. The final score $s_{tot}\left(x_{i,j}\right)$ is calculated as shown below in Equations (10) and (11).

$$p = arg\,max\left(S_Z\left(x_{i,j}\right)\right) \tag{10}$$

$$s_{tot}\left(x_{i,j}\right) = t_{conf\,i,j} + S\left(x_{i,j}\right)_p + S_Z\left(x_{i,j}\right)_p \tag{11}$$

The proposed system outputs the prediction results at the appropriate time by accumulating the total identification score for every tracked face. The score dictionary (with multiple tracking IDs) works as a key by adding the scores from the same identification results and subtracting the scores from different identification results. If the score dictionary value falls below 0, the face ID of the respective tracking ID is changed to the face ID of the last identified embedding. Therefore, the identification results are confirmed even with a relatively low cumulative number when a higher total identification score is assigned to the same face ID, but this can be easily ignored if a low total identification score is assigned to a specific face ID. The identification results are confirmed when the accumulated total identification scores exceed an arbitrary threshold β; thus, the threshold β works as a factor for determining how sensitive the system is in recognizing the data. It is difficult to determine the sensitivity of the system that would ensure the best predictions for all situations happening in the real world. As such, the threshold value with the highest possible prediction accuracy can be determined through an experiment on a specific dataset. A preliminary experiment was conducted to determine the appropriate threshold β and the result is provided in Section IV-C.

The total identification score $s_{tot}\left(x_{i,j}\right)$ for each face images on a single frame is calculated by summing the detection confidence, z-score, and embedding distance score as shown in Equation (11). However, the discrimination effect is low when this value is simply accumulated, as the data are distributed around a specific value. A concentrated distribution can be dispersed by applying activation functions. An activation function that works as an exponentiation of a natural constant only disperses the data concentrated on high values. A sigmoid function has a dispersion effect but only disperses the data positioned at the center, and thus cannot be applied because the data contributing to the total identification score are not necessarily concentrated at the center. Therefore, the activation is performed by dispersing the score distribution through normalization and exponentiation. Equation (12) expands the score distribution from the criteria α via a cubic activation.

$$s_{scaled}\left(x_{i,j}\right) = \left(s_{tot}\left(x_{i,j}\right) - \alpha\right) + 1)^3 + (\alpha - 1) \tag{12}$$

A preliminary experiment was conducted to identify the score value with the most distributions to set it as the mean value, and the result is provided in Section IV-C. The data distribution was adjusted to set the mean value as 1; then, the scores are restored to criteria α to disperse the data

Algorithm 1 Real-Time Criminal Face Recognition

```
INPUT:  Face detector $\mathrm{f_{det}}$, Face encoder $\mathrm{f_e}$,
        Down-sampling ratio $s_w, s_h$,
        Target face image set $Y$
dictionary $\mathcal{T}$
$i \leftarrow 0$
while true do
    $x_i \leftarrow$ GetFrameImage $(i)$
    if $x_i$ is empty then
        /* terminate process */
        break
    end if
    $x_i^* \leftarrow$ DownScale $\left(x_i, \{\frac{1}{s_w}, \frac{1}{s_h}\}\right)$
    $\mathbb{B}_i^* \leftarrow f_{det}\left(x_i^*\right)$
    $\mathbb{B}_i \leftarrow \{\mathbb{b}_{i,j} \odot \{s_w, s_h, s_w, s_h, 1\} | \mathbb{b}_{i,j} \in \mathbb{B}_i\}$
    for $\mathbb{b}_{i,j} \in \mathbb{B}_i$ do
        $x_{i,j} \leftarrow$ Crop$(x_i, \mathbb{b}_{i,j,:-1})$
        /*target face embedding $\mathrm{f_e(y_k)}$ is
        pre-encoded*/
        calculate $S_z\left(x_{i,j}\right)$
        face id $p \leftarrow$ argmax $\left(S_Z\left(x_{i,j}\right)\right)$
        tracking id $t \leftarrow$ Sort$(b_{i,j,:-1})$
        calculate $s_{tot}\left(x_{i,j}\right)$
        $s_{scaled}\left(x_{i,j}\right) \leftarrow \left(s_{tot}\left(x_{i,j}\right) - \alpha\right) + 1)^3$
        if $t \notin \mathcal{T}.keys()$ then
            $\mathcal{T}.item\,(t) \leftarrow \{s_{scaled}\left(x_{i,j}\right), p\}$
        else than
            $c, q \leftarrow \mathcal{T}.item\,(t)$
            if $q = p$ then
                $\mathcal{T}.item\,(t) \leftarrow \{c + s_{scaled}\left(x_{i,j}\right), q\}$
            else if $c - \mathrm{s_{tot}}\left(\mathrm{x_{i,j}}\right) < 0$ then
                $\mathcal{T}.item\,(t) \leftarrow \{s_{scaled}\left(x_{i,j}\right) - c, p\}$
            else then
                $\mathcal{T}.item\,(t) \leftarrow \{c - s_{scaled}\left(x_{i,j}\right), p\}$
            end if
        end if
        if $\mathcal{T}.item\,(t)_0 > \beta$ then
            ReportResult$\left(x_{i,j}, \mathbb{b}_{i,j,:-1}, \mathcal{T}.item\,(t)\right)$
        end if
    end for
    $i \leftarrow i + 1$
end while
```

distribution and output more accurate prediction results more quickly.

The overall score calculation and accumulation steps of the proposed system is summarized in Algorithm 1. The procedure iterates while the video is being input in real-time. Therefore, if a criminal face exceeding the threshold β is recognized, the result is reported to the streaming process with on-memory DB and this iteration continues.

E. PREPROCESSING FOR FACE IDENTIFICATION

Face identification (regardless of technique) entails a unique face alignment process. Face alignment refers to an image processing method for refining images so that parts other than a face do not interfere with the feature extraction from the detected face region. Several studies have been conducted on face alignment techniques for enabling face poses or viewpoints to be identified using facial landmark detection and 3D face restoration techniques based on deep neural networks (which have drastically advanced in recent years). Fundamentally, a precise facial landmark or segmented mask must be acquired, as a face boundary needs to be estimated to perform face alignment. Most face identification processes are trained based on unique face alignment processes; hence, a verified identification performance can be reenacted only if the feature extraction network used for identification is used along with the face identification process applied for training. In the face detection process D of the system configuration proposed in this study, multitasking can be conducted to simultaneously predict both the face bounding box and landmark. However, the verified performance of the relevant techniques cannot be easily reenacted, because the feature extraction network for the identification process, I is not trained based on the landmarks of the detection process D. The multitasking in the detection process D is applied during the training process of the proposed system to improve the detection performance; only the bounding box output is obtained without using landmarks among the detection results, as the computation efficiency is not degraded by the multitasking. The face region images are aligned by separately conducting the unique facial landmark detection of the FaceNet method, which is the basis of the identification process I. There are two methods for outputting landmarks: a 68-coordinate detection method and a five-coordinate detection method. An experiment showed that using five coordinates for detection demonstrated better identification performance while improving the processing speed by eliminating unnecessary operations. When the region required by the face alignment process differs from the face region predicted in the detection process D, normal face alignment is not performed, and the identification performance is degraded. This problem can be resolved by slightly expanding the detected face region before the identification process. The face region expansion is performed by adjusting the width and height of the bounding box at an arbitrary ratio. The minimum expansion ratio for face alignment and identification performance in the proposed system was measured through an experiment, resulting in an expansion ratio of 1.3.

IV. EXPERIMENTAL RESULTS

A series of experiments were performed to measure the effects of the method proposed herein. In particular, the recognition accuracy in actual situations was reenacted by simulating situations where specific targets were captured with minimum latency in an environment where high-resolution videos are input. Each process constituting the system complied with a conventional input/output format, and therefore, various face detection or identification methods could easily be applied. Accordingly, comparison results between the method proposed in this study and major facial recognition methods that are more easily applied are presented below.

A. DATASET FOR EVALUATION

For measuring the effects of the proposed system, a test dataset had to be configured to determine whether the dataset was appropriate for the intended use. In general, the systems for detecting objects in video data are typically evaluated by measuring the classification accuracy for each video. However, the classification accuracy for individual videos cannot fully explain the operation performance in real-life situations, as the process of catching criminals through crime-prevention surveillance cameras must consider how the videos are input regularly over a long period of time. First, sampling must be performed by setting an arbitrary sample size, as the size of T in RGBT data is not limited. This can lead to one of the most serious issues, i.e., being unable to define the answer label when multiple persons are caught in the sampled individual video. The video datasets in [38] and [39] can be used for evaluating real-time facial recognition performance, but they are not appropriate for simulating the task targeted in this study, because details such as the imaging format or face size vary significantly. The majority of datasets are composed of fragmentary and discontinuous images; there are some datasets of videos, but it is difficult to conduct a matching experiment with mug shots captured and processed in a fixed environment. In addition, the resolution of the video data included in such datasets are extremely low or not uniform. As such, an experiment needs to be conducted under fixed input conditions at FHD resolution to verify whether the system can handle real-time processing. Moreover, the appropriate datasets vary depending on the specific purpose, such as face detection, identification, or recognition. The purpose of this study was to examine the facial recognition performance in a situation closely resembling an actual application environment. Data annotated in as detailed a manner as possible were required for evaluating the overall performance. The research team generated a dataset for directly evaluating the system performance in the experiment. Specifically, 17 experiment participants were instructed to move in three different directions, and their movements were captured by two cameras at fixed locations to collect video data.

In addition, the dataset was configured with a total of 105 short FHD videos by adding video data captured in different environments. Regarding the data for evaluating the generalization performance of deep learning-based computer vision systems, it is recommended that the dataset be configured by including environmental factors that may happen in real life. However, certain issues are difficult to solve when using FHD video data; in particular, in this case, the data are created from surveillance footage, such that unspecified individuals may appear unexpectedly in an evaluation dataset. Other issues are that (1) an extremely high labor cost is required for data labeling, and (2) ethical issues are involved in capturing a large number of unspecified individuals. Several large-scale open datasets are already available for the training and evaluation of face detection and identification systems, and therefore, it was not necessary to build a new dataset to prove the generalization performance of each process constituting the overall system. However, a new dataset was constructed for the experiment, as it was conducted in an environment similar to real life. The focus was on experimentally proving the following performances using this dataset:

1) Processing speed of FHD resolution videos
2) Conclusive facial recognition performance through face tracking
3) Measuring the effects of face detection information on identification performance

When a dataset was captured, the viewpoint was moved vertically while the subjects faced left, right, and front to ensure that a certain level of deformation did not occur. In addition, the subjects were instructed to walk forward to measure the correlation between the identification accuracy and the resolution of the bounding boxes of the detected faces.

B. IMPLEMENTATION DETAILS

All of the comparison models and the proposed RTCDS system were implemented using NVIDIA GeForce RTX 3090 GPU in Python 3.8, and PyTorch 1.8.1 for the experiment. The RetinaFace framework was selected as the baseline for the experiment, and certain post-processing operations were vectorized for real-time processing. ResNet50 was mostly used as the feature extraction network for face detection. The face detection region adjustment ratio was empirically set to 1.3. Pre-trained weights provided in the literature and suggested for the respective frameworks or systems were used in all same deep neural networks. In general, the task of identifying faces in real life must guarantee the performance of correctly matching hundreds of detection targets (such as mug shots) images with actual subjects. The front-face captured images of the K-face dataset were used in the experiment, as hundreds of face images playing the role of detection targets were needed in addition to participants of the dataset generation. The individuals to be included in the detection target group were randomly selected by including the face images of the participants captured by the research team in the K-face dataset for fair experimental results.

C. APPEARANCE RECOGNITION MODULE

The experimental results under all conditions of the experimental systems were evaluated using metrics (including the accuracy and F-1 score) by creating a prediction confusion matrix for the ground truth. Table 1 presents the recognition results according to the conditions of the proposed system for the dataset created by our team. The proposed system

TABLE 1. Performance according to scaled total score by thresholds.

Criteria α	threshold β	Acc.	F-1
4.0	10	0.880	0.916
	15	0.885	0.926
	20	0.891	0.942
	25	0.890	0.942
	30	0.872	0.931
4.5	10	0.894	0.940
	15	0.900	0.943
	20	0.872	0.931
5.0	10	0.883	0.925
	15	0.890	0.942
	20	0.853	0.921

tracks face objects and predicts scores at each frame unit, and then outputs results when the identification results based on the accumulated scores for each face object are convincing. Therefore, the overall operational performance of the system is determined by aggregating the scores for each face object unit.

Our preliminary experimental results for obtaining the optimal α and β values are shown in Figure 8. The final result was not determined until a face detected in a captured video disappeared from the video, thereby being aggregated as a false negative. In contrast, a false positive aggregation occurred if the system predicted a different person or object than that corresponding to the detected face. In Table 1, α indicates the reference score used when polarizing scores before the cumulative summing with the total score. As the value of α vincreases, only the prediction results for higher confidence levels are selected for accumulation, resulting in higher precision.

As the results are output and aggregated if the accumulated score exceeds β, the threshold β is also directly related to the precision. These two parameters can be adjusted depending on the size of the identification target list or identification method.

D. COMPARISON OF PROPOSED INSTANCE-LEVEL METHOD WITH EXISTING FRAME-LEVEL METHODS

In this paper, we proposed the tracked instance-level face identification method which generates and tracks an instance for a detected face and then outputs cumulative identification result. This method is very suitable for video input data in comparison to existing frame-level identification methods including FaceNet [33], SphereFace [40], CosFace [41], and ArcFace [42]. We performed experiments to compare the performance of the proposed method with existing frame-level methods on our video dataset which captured a pedestrian approaching the camera. Each video is shot from various angles from which the front of the face can be seen. Our dataset consists of 105 videos shot at 30 FPS for approximately 10 s, with each video featuring only one person.

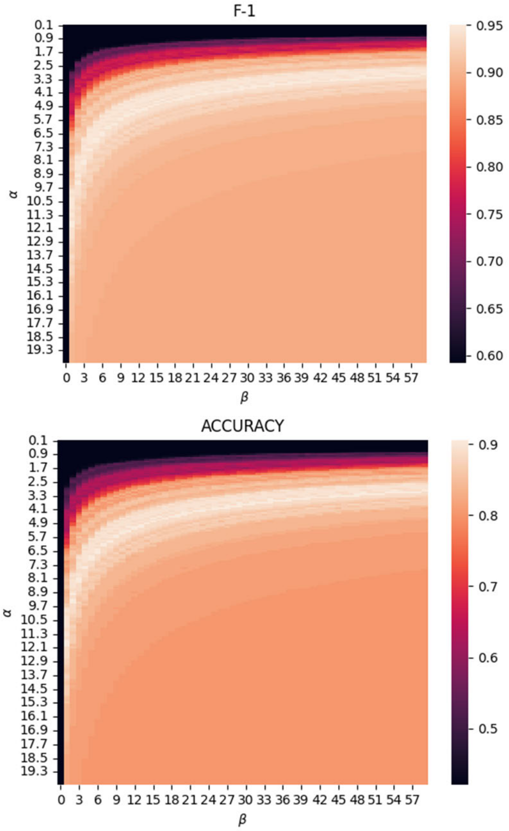


FIGURE 8. Heatmap of accuracy and f1-score according to α, β.

For real-time detection, we aim to minimize the trade-off between performance and speed in the existing RetinaFace-based face detection process. Many face detectors that deal with the scale variation problem adopt FPN. In our work, we also needed a Scale Invariant Detector, and many detectors do not meet real-time processing requirements. RetinaFace can perform real-time processing on a single GPU while adopting the ResNet 50 network as the backbone. Strictly speaking, the standard for real-time processing may vary depending on the FLOPS of the GPU model or specific conditions, making it difficult to objectively compare processing speed and performance with other methods. We ensured sufficient real-time processing speed and performance while fixing detection conditions and presented the degree of performance improvement using the tracking and score accumulation method proposed in our experimental results.

The first experiment is to evaluate the performance of existing frame-level methods when they are directly applied to our video dataset. Table 2 shows both the average accuracy and average F-1 score measured per frame of the video dataset based on the prediction result at the frame level. All the results are insufficient because the existing methods were designed for image/frame-level input, not for video input.

On the other hand, the second experiment is to evaluate the performance of the proposed instance-level method by using

TABLE 2. Detection and identification performance evaluations per frame.

Identification Method	Details	Acc.	F-1
FaceNet[33]	-	0.362	0.499
SphereFace[40]	Single model	0.412	0.522
SphereFace[40]	Three-path ensemble	0.418	0.543
CosFace[41]	Support vector (SV)-additive margin (AM)-SoftMax	0.439	0.595
ArcFace[42]	MS1MV2 + R100 + R	0.573	0.631

TABLE 3. Detection and identification system performance evaluations per video.

Identification Method	Details	Acc.	F-1
FaceNet[33]	-	0.900	0.947
SphereFace[40]	Single model	0.907	0.949
SphereFace[40]	Three-path ensemble	0.914	0.961
CosFace[41]	SV-AM-SoftMax	0.910	0.955
ArcFace[42]	MS1MV2 + R100 + R	0.921	0.966

TABLE 4. Comparison of performance between frame-level and tracking-level prediction.

Identification Method	Acc.	R	P	F-1
Frame-Level	0.329	0.314	0.324	0.319
Tracking-Level (Ours)	0.788	0.789	0.82	0.804

existing methods as a tool for frame-level face detection and identification. Table 3 shows both the accuracy and F-1 score of the proposed method when using each existing method, i.e., FaceNet, SphereFace, and CosFace as a tool. The evaluation results show a noticeable performance improvement in the proposed tracked instance-level face identification method, in comparison to those of Table 2. By accumulating the information about a person appearing in a series of frames in each tracking instance, the confusion about the prediction occurring in each frame is eliminated. This high accuracy and F-1 score enable the proposed method to be applied to tasks requiring real-time recognition of a person appearing in a video.

We made some modifications to our dataset in order to provide a more objective performance evaluation. Specifically, the dataset used for the performance evaluation in Tables 2 and 3 does not include false samples in the ground truth. As a result, measuring Precision is meaningless. To more accurately compare performance, we experimentally assigned false labels to half of the existing true data labels, in order to see how well the faces are distinguished. The results of this experiment are shown in Table 4. Method in Table 4 used the same FaceNet-based method as listed in Tables 2 and 3. Although the absolute performance values, including accuracy, have decreased, the performance is still superior to frame-based prediction methods. Figure 9 shows a visualization of the performance evaluation results in Table 4 as a Confusion matrix. The numbers written in each cell of the confusion matrix represent the number of predicted unit samples in the dataset.

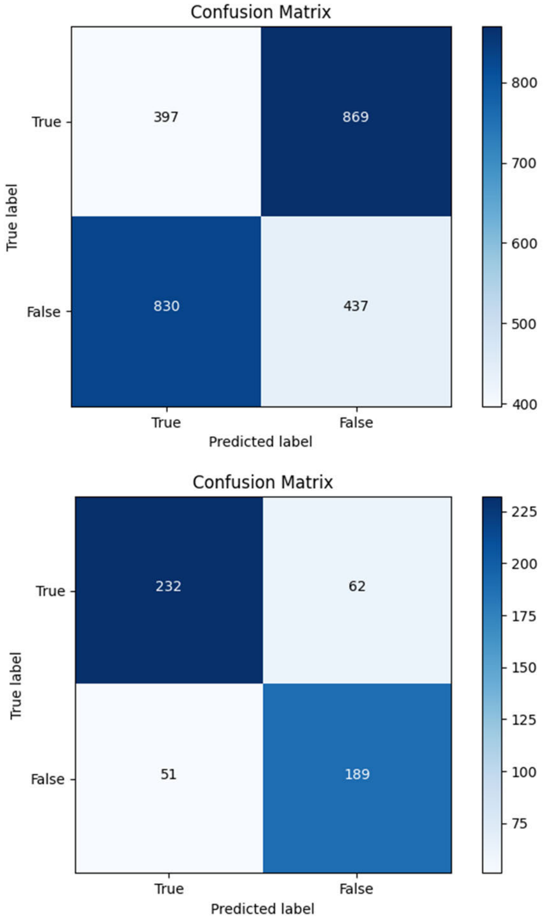
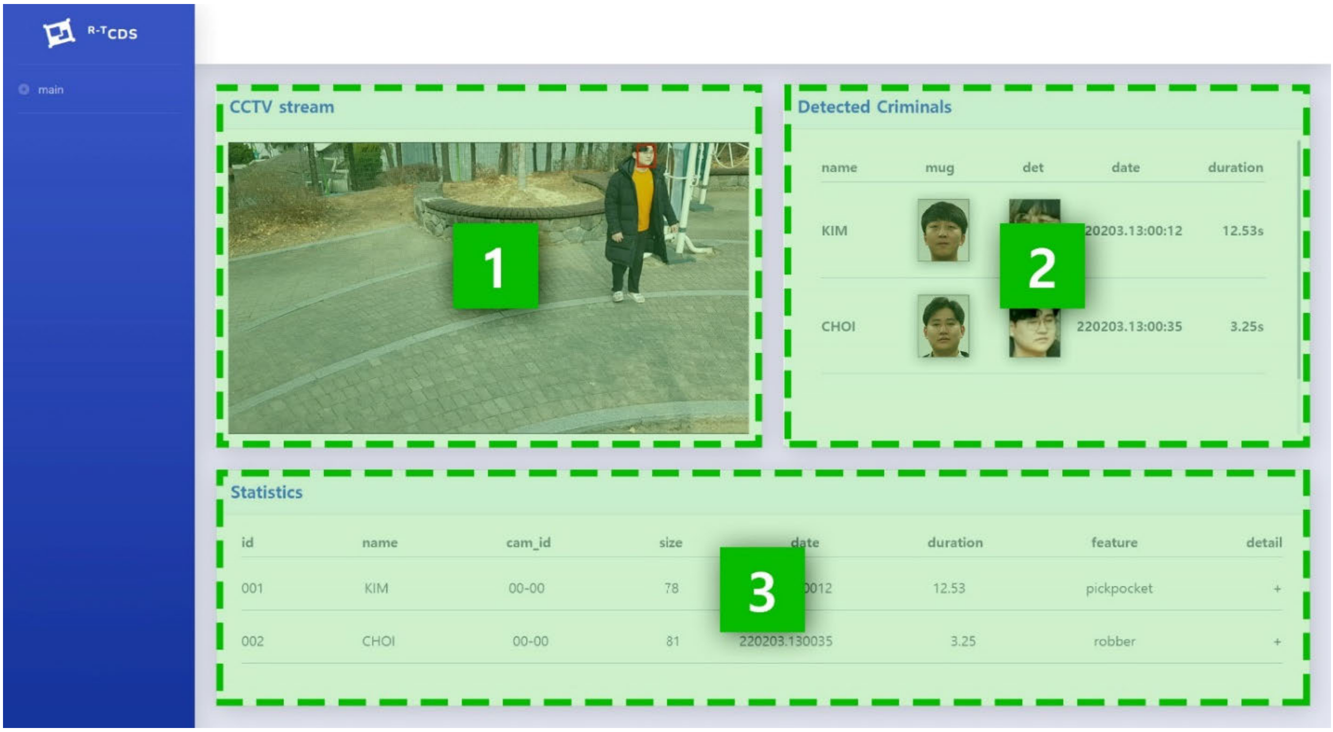


FIGURE 9. Confusion matrix of evaluation result shown in Table 4.

E. IMPLEMENTATION OF THE PROTOTYPE APPLICATION

The prototype shown in Figure 10 was implemented and tested to verify and confirm the operation of the model, as well as the platform proposed in this study. The prototype included a server computer equipped with a deep learning model connected to a surveillance camera, web server, and web page for an administrator.

The prototype could inspect the images of the currently connected surveillance camera, as shown in Figure 10-(1). The mug shot of the face of a detected criminal in a certain section and the detected time are shown in Figure 10-(2).



name	mug	det	date	duration
KIM			20203.13:00:12	12.53s
CHOI			220203.13:00:35	3.25s

id	name	cam_id	size	date	duration	feature	detail
001	KIM	00-00	78	0012	12.53	pickpocket	+
002	CHOI	00-00	81	220203.130035	3.25	robber	+

FIGURE 10. Web execution screen of the proposed system prototype.

Detailed information, such as the address and personal details of the detected criminal, are shown in Figure 10-(3). Furthermore, the prototype includes a push notification feature to notify the relevant institutions and personnel of the detected criminal faces.

V. CONCLUSION

Currently, there are too many surveillance cameras for the dedicated staff of the Safety Management Centre to visually check. As a result, crimes are not immediately identified, and much of the recorded footage is discarded without being put to any useful use. Existing studies using surveillance cameras have shown that the awareness of the presence of cameras can induce impulsive criminal behavior, but it is difficult to show a direct preventive effect. There are also limitations in analyzing the recorded images through artificial intelligence, such as not being able to use them for post-mortem investigations or performing real-time identification. This paper proposes a system that detects the appearance of high-risk individuals in real-time based on surveillance video analysis of installed surveillance cameras and expansion of recorded information and includes an application that notifies identification information to public safety agencies through a data pipeline and confirms the person's identity information in real-time.

The proposed system analyzes video footage captured by surveillance cameras in real-time. By using a method that iteratively detects and identifies faces in each frame, the footage can be analyzed immediately without storing it. By proposing a face recognition method that uses down-sampling to identify face positions and utilizes them in the original quality image, the performance of face detection and identification can be improved on the same hardware to enable real-time detection. It contributes to improving the precision of object tracking by storing the location of the detected face in the video and the identification information predicted by the system. The face tracking ID unit also compensates for the problems of the prediction unit when performing face recognition in video data. The face tracking ID unit minimizes the prediction flipping problem caused by the congested embedding problem due to the large size of the embedding DB through the identification score accumulation method. The threshold value used in the identification score accumulation method was detected through experiments to find the optimal threshold. In addition, a dataset was created for evaluation and measurement in the overall experiment. Since the proposed system uses the input and output formats of common face detection and identification systems, it ensures freedom of tuning, which allows practical users to easily apply different models suitable for specific domains. This is evidenced by the improvement in accuracy and F-1 score during migration in the experiments according to the identification method in Tables 2 and 3. In addition, two parameters can be utilized to derive the final score for the tracked object, allowing a high precision or recall to be selected. In our experiments, we obtained an accuracy of 0.900 and an F-1 score of 0.943 for ($\alpha = 4.5$, $\beta = 15$).

The proposed system is a security monitoring system for detecting certain categories of people by analyzing the videos recorded by video security devices such as surveillance cameras in real-time without missing anything through deep learning. This effectively complements the blind spots of the existing visual monitoring system by detecting and recognizing criminals in crime-prone areas and notifying relevant agencies and monitoring personnel in real-time. The main contribution of the proposed system is that it shows higher performance on the same performance hardware through downsampling of the shooting video to ensure real-time processing. In addition, high identification and recognition rates are achieved through the face tracking ID unit and the identification score accumulation method. We expect that the proposed system will complement the blind spots of the existing surveillance system, leading to rapid response for incident and accident prevention through national projects such as identification of criminal access, location search for missing children and criminals, and protection of national facilities.

REFERENCES

[1] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems-current status and future trends," *Comput. Electr. Eng.*, vol. 70, pp. 736–753, Aug. 2018.

[2] J. T. Pickett, C. Mancini, and D. P. Mears, "Vulnerable victims, monstrous offenders, and unmanageable risk: Explaining public opinion on the social control of sex crime," *Criminology*, vol. 51, no. 3, pp. 729–759, Aug. 2013.

[3] D. M. Button, M. DeMichele, and B. K. Payne, "Using electronic monitoring to supervise sex offenders: Legislative patterns and implications for community corrections officers," *Criminal Justice Policy Rev.*, vol. 20, no. 4, pp. 414–436, Dec. 2009.

[4] M. C. Hoffman, "Smart cities: A review of the most recent literature," *Informatization Policy*, vol. 27, no. 1, pp. 3–35, 2020.

[5] B. C. Welsh and D. P. Farrington, "Public area CCTV and crime prevention: An updated systematic review and meta-analysis," *Justice Quart.*, vol. 26, no. 4, pp. 716–745, Dec. 2009.

[6] E. L. Piza, B. C. Welsh, D. P. Farrington, and A. L. Thomas, "CCTV surveillance for crime prevention: A 40-year systematic review with meta-analysis," *Criminol. Public Policy*, vol. 18, no. 1, pp. 135–159, Feb. 2019.

[7] E. L. Piza, "The crime prevention effect of CCTV in public places: A propensity score analysis," *J. Crime Justice*, vol. 41, no. 1, pp. 14–30, Jan. 2018.

[8] G. Alexandrie, "Surveillance cameras and crime: A review of randomized and natural experiments," *J. Scandin. Stud. Criminology Crime Prevention*, vol. 18, no. 2, pp. 210–222, Jul. 2017.

[9] R. V. Clarke, "Situational crime prevention," *Crime Justice*, vol. 19, pp. 91–150, 1995.

[10] J. Ratcliffe, *Video Surveillance of Public Places*. Washington, DC, USA: U.S. Department of Justice, Office of Community Oriented Policing Services, 2006.

[11] J.-H. Jeon and S.-R. Jeong, "Designing a crime-prevention system by converging big data and IoT," *J. Internet Comput. Services*, vol. 17, no. 3, pp. 115–128, Jun. 2016.

[12] P. Chen, H. Yuan, and X. Shu, "Forecasting crime using the ARIMA model," in *Proc. 5th Int. Conf. Fuzzy Syst. Knowl. Discovery*, Oct. 2008, pp. 627–630, doi: 10.1109/FSKD.2008.222.

[13] Lubna, N. Mufti, and S. A. A. Shah, "Automatic number plate recognition: A detailed survey of relevant algorithms," *Sensors*, vol. 21, no. 9, p. 3028, Apr. 2021.

[14] K. B. Kwan-Loo, J. C. Ortíz-Bayliss, S. E. Conant-Pablos, H. Terashima-Marín, and P. Rad, "Detection of violent behavior using neural networks and pose estimation," *IEEE Access*, vol. 10, pp. 86339–86352, 2022.

[15] N. A. Abdullah, M. J. Saidi, N. H. A. Rahman, C. C. Wen, and I. R. A. Hamid, "Face recognition for criminal identification: An implementation of principal component analysis for face recognition," in *Proc. AIP Conf.*, 2017, Art. no. 020002.

[16] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, vol. 20, no. 2, p. 342, Jan. 2020.

[17] N. Dagnes, E. Vezzetti, F. Marcolin, and S. Tornincasa, "Occlusion detection and restoration techniques for 3D face recognition: A literature review," *Mach. Vis. Appl.*, vol. 29, no. 5, pp. 789–813, Jul. 2018.

[18] M. Adil, S. Mamoon, A. Zakir, M. A. Manzoor, and Z. Lian, "Multi scale-adaptive super-resolution person re-identification using GAN," *IEEE Access*, vol. 8, pp. 177351–177362, 2020.

[19] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 471–478, doi: 10.1109/SIBGRAPI.2018.00067.

[20] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu, "Partial FC: Training 10 million identities on a single machine," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1445–1449.

[21] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "TinaFace: Strong but simple baseline for face detection," 2020, *arXiv:2011.13183*.

[22] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2001, pp. 1–11.

[23] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[24] L.-F. Chen, H.-Y.-M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, no. 10, pp. 1713–1726, Oct. 2000.

[25] E. Ben-Baruch, M. Karklinsky, Y. Biton, A. Ben-Cohen, H. Lawen, and N. Zamir, "It's all in the head: Representation knowledge distillation through classifier sharing," 2022, *arXiv:2201.06945*.

[26] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi, "GhostFaceNets: Lightweight face recognition model from cheap operations," *IEEE Access*, vol. 11, pp. 35429–35446, 2023, doi: 10.1109/ACCESS.2023.3266068.

[27] Y. Zheng, J. Chang, Z. Zheng, and Z. Wang, "3D face reconstruction from stereo: A model based approach," in *Proc. IEEE Int. Conf. Image Process.*, 2007.

[28] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 534–551.

[29] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.

[30] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

[31] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[32] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[33] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[34] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Comput. Vis. Image Understand.*, vol. 138, pp. 1–24, Sep. 2015.

[35] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[36] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[37] D. King. (2017). *High Quality Face Recognition With Deep Metric Learning*. [Online]. Available: http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html

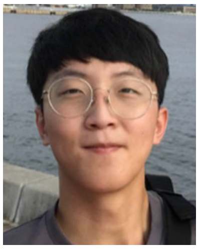
[38] C. Ferrari, S. Berretti, and A. Del Bimbo, "Extended YouTube faces: A dataset for heterogeneous open-set face identification," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3408–3413.

[39] Y. Lin, S. Cheng, J. Shen, and M. Pantic, "MobiFace: A novel dataset for mobile face tracking in the wild," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.

[40] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6738–6746.

[41] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[42] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.



HYUN-BIN KIM received the B.S. degree from the Department of Computer Science, Kyonggi University, Suwon, South Korea, in 2021, where he is currently pursuing the M.S. degree in computer science with the Department of Computer Science. He was a Research Associate with the Computer Graphics and Image Processing Laboratory, Kyonggi University. His research interests include computer vision, anomaly detection, video processing and enhancement, and medical image analysis.

NAKHOON CHOI received the B.S. and M.S. degrees from the Department of Computer Science, Kyonggi University, South Korea, in 2020 and 2022, respectively, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. He was a Research Associate with the Information Security Laboratory, Kyonggi University. His research interests include blockchain, deepfake detection, information security, and copyright protection.

HYE-JEONG KWON received the B.S. degree from the Division of Computer Science and Engineering, Kyonggi University, Suwon, South Korea, in 2021, where she is currently pursuing the master's degree with the Department of Computer Science. She was a Researcher with the Data Mining Laboratory, Kyonggi University. Her research interests include data mining, artificial intelligence, information systems, emerging health risk mining, and medical data analysis.

HEEYOUL KIM received the B.E., M.S., and Ph.D. degrees in computer science from KAIST, South Korea, in 2000, 2002, and 2007, respectively. From 2007 to 2008, he was with the Samsung Electronics as a Senior Engineer. Since 2009, he has been a Faculty Member with the Department of Computer Science, Kyonggi University. His major research interests include cryptography, security, and blockchain.