A Research on the Application of Face Recognition Algorithm Based on Convolutional Model and Transformer Model in Community Monitoring System

Tan Heyi[1], Byung-Won Min[2*]
[1]Ph.D. Student, Division of Information and Communication Convergence Engineering, Mokwon University
[2]Professor, Division of Information and Communication Convergence Engineering, Mokwon University

컨벌루션 모델과 트랜스포머 모델을 이용한 얼굴식별 알고리즘의 지역 사회 모니터링시스템 적용에 관한 연구

담하의[1], 민병원[2*]
[1]목원대학교 정보통신융합공학부 박사과정, [2]목원대학교 정보통신융합공학부 교수

Abstract The promotion of intelligent security community construction has greatly enhanced the intelligence and safety of residential areas. In order to further establish a security-oriented community, this paper proposes the utilization of facial recognition based on community surveillance footage to identify suspicious individuals. To address the difficulties in capturing facial images caused by factors such as low pixel resolution and varying shooting angles in surveillance footage, the following optimization strategies are proposed in this paper : Firstly, a lightweight global search facial detection network is designed based on convolutional modules and Vision Transformer modules. The Vision Transformer module is introduced to enhance the global retrieval capability of the network. Secondly, the structure of the Vision Transformer module is optimized by adding pooling layers in the feature block extraction and segmentation stage to reduce the number of module parameters. The feature blocks are mapped and computed with the feature maps to improve the corresponding feature correlation. Thirdly, in the face alignment stage, an Anchor Free mechanism is adopted to generate elliptical face localization regions for more accurate fitting of faces and reducing interference from other background information in the final identity recognition stage. Finally, the similarity between faces is calculated using Euclidean space distance to determine corresponding personnel identities. Through relevant experiments and tests on the self-built facial identity dataset in this paper's residential surveillance system, the proposed facial detection network achieves an average improvement of 3.11% in detection accuracy compared to other detection networks, reaching 97.19%. In terms of facial identity recognition, the designed model achieves an average improvement of 3.43% with a recognition accuracy of 95.84.

Key Words : facial recognition, Vision Transformer, ultra-lightweight.

요 약 스마트 지역사회 건설이 추진됨에 따라 지역사회의 지능화와 안전성이 크게 향상되었으며, 스마트 지역사회 보안을 더욱 구축하기 위해 본 연구에서는 지역사회 모니터링 비디오 화면을 기반으로 얼굴 식별을 구현하여 의심스러운 사람들의 출현을 보다 정확하게 식별하고 경고하는 것이다. 모니터링되는 영상 화면의 낮은 픽셀과 촬영 각도 변화와 같은 요인으로 인해 얼굴 화면을 캡처하는 데 어려움이 있음을 고려하여 최적화 해결방법을 제안한다. 첫째, 컨벌루션 모듈과 비전 트랜스포머 모듈을 결합한 경량화적인 얼굴 감지 네트워크를 전체적으로 검색하는 것을 설계했다. 네트워크의 전체 검색 능력을 향상시키기 위해 Vision Transformer 모듈을 네트워크에 처음으로 추가했다. 둘째, Vision Transformer 모듈에 대해 심층 구조 최적화를 수행했는데, 특징 블록 추출 및 분할 단계에서 모듈의 매개변수 수를 줄이기 위해 풀링 레이어가 추가되었다. 동시에 특성 블록과 특성 맵을 일치시켜 특성 간의 연관성을 향상시킨다. 셋째, 얼굴 정렬 단계에서 Anchor Free 메커니즘을 기반으로 타원형 얼굴 위치 영역을 설계했다. 이렇게 하면 얼굴과 더 정확하게 일치시키고 배경 정보의 간섭을 줄이며 최종 식별을 더 정확하게 할 수 있다. 넷째, 유클리드 공간 거리를 사용하여 얼굴의 유사성을 계산하여 사람의 신분을 결정한다. 관련 실험 테스트 후, 설계한 얼굴 감지 네트워크는 감지 정확도가 평균 3.11% 향상되어 97.19%에 도달했다. 얼굴 식별 측면에서 모델 인식 정확도는 평균 3.43% 향상되어 95.84%에 도달했다.

주제어 : 얼굴식별, 비전 트랜스포머, 울트라 라이트웨이트

*교신저자 : 민병원(minfam@mokwon.ac.kr)
접수일 2023년 7월 15일 수정일 2023년 8월 24일 심사완료일 2023년 8월 28일

1. Introduction

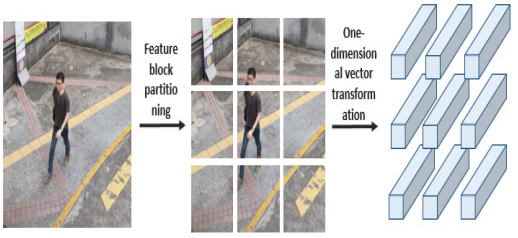
The security surveillance system is widely applied in various locations, providing a user-friendly solution that simplifies the work of security personnel. By monitoring the surveillance footage, they can effectively fulfill their patrol duties. However, the reliance on manual video analysis poses challenges as the workload increases. This can lead to decreased stability and effectiveness over time. To address this issue, this paper proposes leveraging computer vision technology to identify individuals in the surveillance footage of residential communities. By harnessing the stable performance of computers, this approach aims to enhance the security and monitoring capabilities of the community.

Currently, there are several solutions that integrate surveillance technology with computer vision to achieve intelligent analysis of surveillance footage. For example, Wang Xin and Liu Xiaolan proposed an improved human pose feature extraction network that combines the serial frame merging module and pose residual module to simplify the feature map scale of object detection. This enables the detection and analysis of worker behavior in surveillance images [1]. Zhang Dexiang and Wang Jun introduced an attention mechanism-based multiscale full-scene surveillance object detection network for complex urban monitoring scenarios. This network, based on the Yolov5s model, incorporates an attention mechanism feature extraction module to enhance target features and suppress background features, thereby improving the feature extraction capability [2].In terms of facial detection and recognition, Xia Guishu and Zhu Zihan proposed an improved Swin Transformer facial recognition model. By introducing a multi-level feature fusion module, the model enhances the representation of facial features to improve recognition accuracy [3]. Ji Ruirui and Xie Yuhui developed a Vision Transformer-based facial recognition method. They utilized the Shuffle Transformer as the backbone network for feature extraction, capturing global information of feature maps through self-attention mechanisms and Shuffle operations to establish long-distance dependencies between feature points, thereby improving the model's feature perception ability [4]. In summary, there are currently mature solutions for detecting and analyzing surveillance videos and recognizing facial identities. This paper aims to utilize surveillance footage for facial identity detection and recognition in order to enhance the intelligence of the surveillance system. However, there are some challenges in this task. Firstly, the size and appearance of the targets in the footage vary due to the different distances between the targets and the surveillance cameras. Secondly, unlike conventional facial recognition, the facial regions in surveillance footage occupy a smaller number of pixels, which makes it difficult to analyze and compare facial features. Therefore, the focus of this paper is to design a facial detection network that can enhance feature extraction capabilities and increase facial feature information, thereby improving subsequent

1.1 Related technologies

In the research on face recognition technology, the performance of the network is enhanced by incorporating the Vision Transformer module. The Vision Transformer module is derived from the optimization of the Transformer module in the field of natural language processing. The Transformer module can compute attentions for one-dimensional input data, enabling long-range feature dependencies. In the context of computer vision, this feature manifests as the extraction of global features, complementing the local feature extraction capability of convolution modules. Therefore, the Vision Transformer module is increasingly being integrated into convolutional networks to improve network

information [5]. However, since the input data in computer vision tasks is two-dimensional, it cannot be directly inputted into the Transformer module for global attention calculation. Therefore, preprocessing and transformation operations need to be performed on the feature data. In the Vision Transformer module, the input feature map is first divided into equally sized feature blocks. Then, the feature blocks are converted into one-dimensional arrays with sequential characteristics using one-dimensional vectors, referred to as tokens. This process is illustrated in Figure.



[Fig. 1] Preprocessing Operations in the Vision Transformer Module

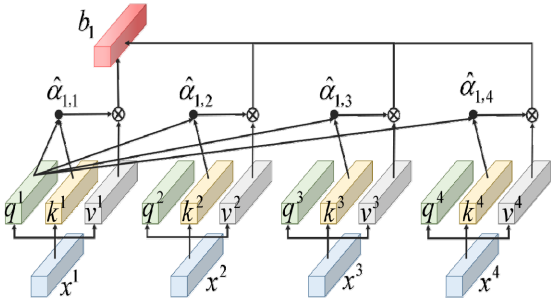
Later, the calculation process of the Transformer module is applied to each token for self-attention calculation [6]. As shown in Figure 2, let's assume that the transformed tokens of the feature blocks are represented as x^1, x^2, x^3, x^4. To calculate the feature relationships between tokens, each x^i undergoes further feature extraction using three vectors: q (query), k (key), and v (value). The q vector is used to query the relevance between the current token and other tokens. The k vector represents the relevance of the current token to other tokens when their q vectors are queried. The v vector represents the feature representation of the token. Taking x^1 as an example, when extracting the relationship features with other input elements, its q^1 is queried with k^1, k^2, k^3, k^4, resulting in a set of weight vectors [α^_1,1, α^_1,2, α^_1,3, α^_1,4]. Then, each weight vector is combined with the input features to obtain the final relationship expression output b^1 between token x^1 and other tokens. In actual calculations, self-attention calculation is performed simultaneously for all tokens, reflecting the multi-head self-attention calculation feature of the Transformer [7]. The self-attention :

$$Attention(Q,K_i,V_i) = soft\max(\frac{Q^T K_i}{\sqrt{d_k}})V_i$$

In the above equation, d_k represents the dimension of the vectors. Due to the potential dimensionality explosion when multiplying vector q and vector k, it is necessary to divide by $\sqrt{d_k}$ to reduce the dimensionality [8]. The further calculation of multi-head self-attention can be expressed as follows:

$$MultiHead(Q,K,V) = \sum(head_1, head_2, ..., head_n)W^0$$

$$head_i = Attention(Q_i, K_i^j, V_i^j)$$



[Fig. 2] Vision Transformer Self Attention Calculation

2. Design of Face Detection and Recognition Network

For the design of a face detection and recognition network for community surveillance, this paper divides the network functionality into two stages. In the first stage, the network

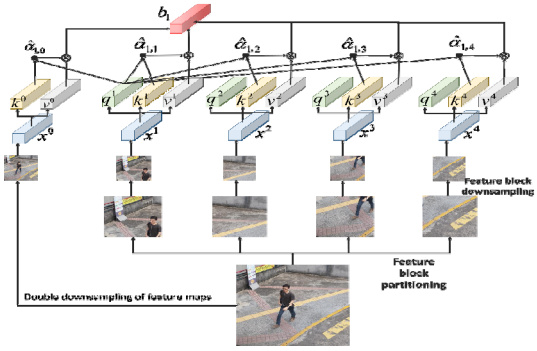
extracts the facial regions from the video frames and performs face alignment. In the second stage, it uses Euclidean distance in the feature space to compare the detected facial features with the database to determine identification matches.

2.1 Face detection network

In the design of network for detection, this paper proposes combining the Vision Transformer module with the convolutional module to construct a feature extraction network. The Vision Transformer module is used to globally extract features from the input feature map in order to obtain relevant positional features of the target. Then, the convolutional module is utilized to locally extract features from the facial region, obtaining finer facial features for landmark calculation.

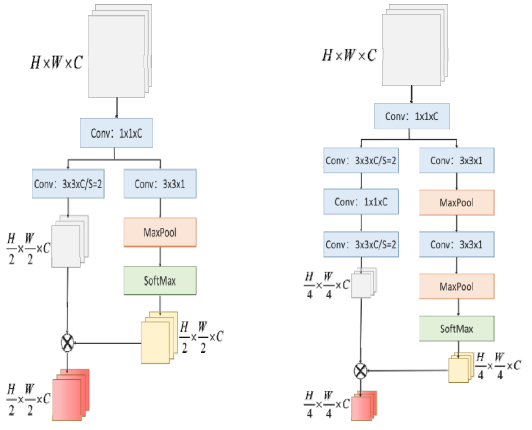
2.1.1 Lightweight Transformer-Lite Module Design

As mentioned in the first section, the Vision Transformer module performs global feature extraction on the feature map by dividing it into feature blocks and conducting multi-head attention calculations to capture the relationships between the blocks. However, this operation has two drawbacks. Firstly, dividing the feature map into blocks partially disrupts the integrity of the features. Secondly, when a large number of feature blocks are created, more computational resources are required for global attention calculations. Moreover, considering the application scenario of the network, quick identification of human targets is required, and in surveillance footage, the facial region only occupies a small pixel area. Therefore, when using the standard Vision Transformer module for calculations, it is necessary to divide the feature map into small enough blocks to ensure global calculations of relevant features, which leads to redundant computational waste. Taking into account the limitations of the Vision Transformer module itself and the characteristics of the application scenario, this paper proposes optimizations for the standard Vision Transformer module, designing a Transformer-Lite with token simplification and global mapping in its multi-head self-attention mechanism. The overall structure of the Transformer-Lite is shown in :



[Fig. 3] Transformer-Lite Module Structure

The operation of reducing computational complexity in the Transformer-Lite module begins with the preprocessing stage of converting the feature map into tokens. Taking the diagram above as an example, the feature map is first divided into four feature blocks. Then, combining downsampling and spatial attention calculations, the feature blocks are reduced in size to improve the granularity of the features. At the same time, the feature map undergoes downsampling by a factor of two to obtain a feature map with the same size as the feature blocks. In this stage, the paper first divides the feature map into coarse-grained feature blocks. Then, through attention mechanisms, key features are extracted and the granularity is enhanced to obtain fine-grained feature blocks that can be converted into tokens. The process :



(a) Feature block fine-grained conversion

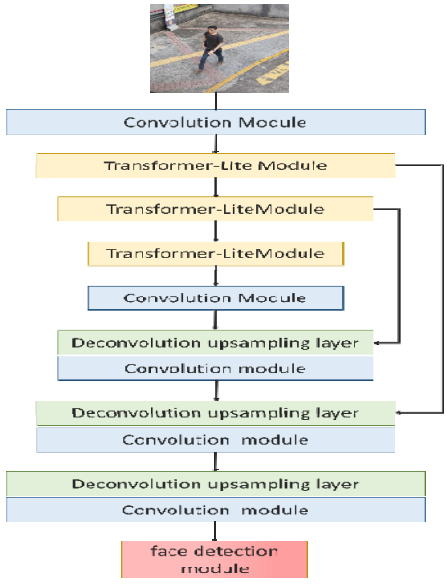
(b) Fine-grained transformation of feature maps

[Fig. 4] Feature Block and Fine Grain Conversion of Feature Map

As shown in Figure 4, taking the transformation of the feature block as an example, it is first subjected to a 1x1xC convolutional layer for a balanced transition of the input feature information. Then, a dual-branch approach is used for downsampling and enhancing key features. In the first branch, a 3x3xC convolutional layer with a sliding stride of 2 is employed for secondary feature extraction and downsampling, resulting in a smaller-sized feature block. In the second branch, a 3x3x1 convolutional layer is used for channel compression, followed by a second spatial feature extraction. Then, a max-pooling operation is performed on the feature block to further reduce its size, and the spatial weight is extracted using the softmax function. Finally, the spatial weight is assigned to the output of the first branch to weight the target features in the fine-grained feature block. Similarly, the secondary downsampling of the feature map and the downsampling of the feature blocks are achieved by using two convolutional downsampling operations and max-pooling operations in the two branches, respectively. In summary, by combining the spatial attention mechanism to enhance the importance of target features and map them to finer-grained feature blocks, the computational complexity is reduced while ensuring that the feature information is not lost. In the subsequent multi-head attention calculation stage, as shown in Figure 3, when each token undergoes self-attention calculation, it is simultaneously mapped with token 0 obtained from the transformation of the feature map, resulting in feature correlation information between the current feature block and the global feature map. This avoids the issue of feature loss caused by feature map segmentation.

2.1.2 Design of detection network based on encoding and decoding architecture

In the face detection network, a feature extraction network with an encoder-decoder structure was designed to extract relevant facial features from surveillance footage. The network structure is illustrated in Figure 5 below:

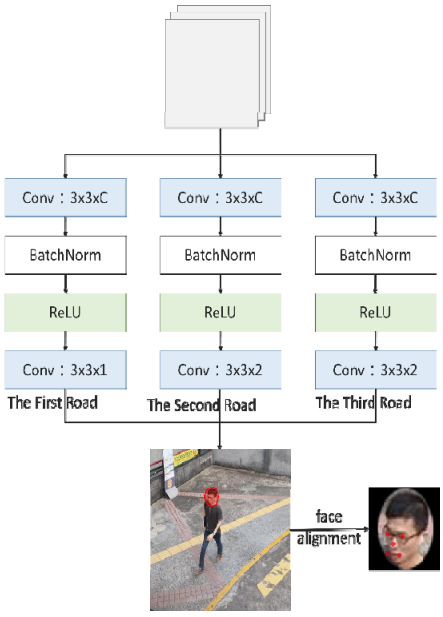


[Fig. 5] Feature Extraction Network Structure

The overall network structure consists of an encoder and a decoder. In the encoder network structure, the paper primarily utilizes the

Transformer-Lite module to extract global features. At the input of the encoder, the surveillance frame footage is first processed by a convolutional module to increase the depth from RGB channels to 64 channels. Then, three Transformer-Lite modules are employed for downsampling and encoding operations. In the encoding stage, the Transformer module is relied upon to perceive the global features, capture the correlation between the face region and other regions, and determine their respective positions in the feature map. Subsequently, the obtained feature map is inputted into the decoder, which consists of four convolutional modules connected through upsampling. The upsampling convolutional operations in the decoder allow the facial features to be progressively magnified for better extraction, benefiting subsequent face detection and alignment operations. Additionally, for the overall design of the feature extraction network, the paper maps the output of the Transformer-Lite modules in the encoder to the feature map in the decoder with the same dimensions, thereby supplementing the global feature information.

For facial detection in residential surveillance systems, where there are typically no densely crowded areas in the community, only a small number of individuals appear in a single monitoring frame. Therefore, using the Anchor mechanism for detection leads to excessive redundant computation, which is not conducive to real-time detection. Hence, in this paper, we propose the use of the Anchor-Free mechanism to reduce computational redundancy and introduce an elliptical face region detection to better capture facial features and avoid interference from other factors. The designed detection structure is illustrated in the following diagram:



[Fig. 6] Anchor Free detector design

As shown in Figure 6, there are a total of three detection branches in the detector module, with the first branch responsible for detecting the center point of the elliptical region; The second branch is responsible for fine-tuning the coordinates of the center point; The third branch is responsible for predicting the length of the major and minor axes of the ellipse. In the first branch, the feature map parameters output through convolution calculation are represented as, because the model only detects targets such as "faces", the number of channels is 1. At the same time, the range of values for each unit in its output feature map is [0,1]. When the feature unit is located in the central store area, its value is 1, and vice versa, it is 0. For further prediction of the center point of an elliptical region, this article uses Gaussian kernels for calculation, with the formula:

$$Y_{xy} = \exp(-\frac{(x-x_{Truth})^2+(y-y_{Truth})^2}{2\sigma_p^2})$$

In the above equation, x_{Truth} and y_{Truth} represent the actual coordinates, while x and y represent the predicted coordinates by the model. σ_p represents the standard deviation of the Gaussian kernel. The output Y_{xy} represents the deviation between the predicted position and the actual position. When Y_{xy} approaches 1, it indicates that the predicted center point coordinates are close to the actual coordinates, while a value further from 1 indicates a larger deviation. The loss calculation for branch one can be expressed as:

$$Loss_{1st} = -\frac{1}{N}\sum_{xy}\begin{cases}(1-\hat{Y}_{xy})^{\alpha}\log(\hat{Y}_{xy}), & Y_{xyc=1}\\ (1-Y_{xy})^{\beta}(\hat{Y}_{xy})^{\alpha}\log(1-\hat{Y}_{xy}), & otherwise\end{cases}$$

In the given equation, N represents the number of predicted center points, $\hat{Y}_{xyc}$ represents the probability that the coordinates of x and y are the center points, α and β are the weights. In the second branch, the parameters of the output feature map are denoted as $H \times W \times 2$. Since the coordinates of the center points are represented by x and y, the number of channels is 2. The calculation of the loss for adjusting the coordinates of the predicted center points can be represented as:

$$Loss_{2nd} = \frac{1}{N}\sum_{p}\left|(p-\tilde{p})\right|$$

In the given equation, p represents the coordinates of the actual center points, and $\tilde{p}$ represents the coordinates of the predicted center points. Finally, the parameters of the feature map output by the third branch are denoted as $H \times W \times 2$, with two dimensions representing the major and minor axes of the elliptical anchor. The loss for the predicted data from this branch can be represented as follows:

$$Loss_{3rd} = \frac{1}{N}\sum_{k=1}^{N}\left|(l-\tilde{l})+(s-\tilde{s})\right|$$

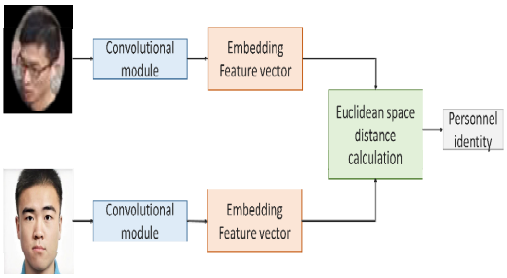
In the given equation, l and s represent the actual lengths of the major and minor axes, while $\tilde{l}$ and $\tilde{s}$ represent the predicted lengths of the major and minor axes. In conclusion, the overall loss for this detector can be represented as:

$$Loss = Loss_{1st} + \lambda Loss_{2nd} + \beta Loss_{3th}$$

In the given equation, λ and β are harmonic weights. As shown in Figure 6, after obtaining the elliptical facial region, the bounding rectangle of the ellipse is extracted. The feature values outside the elliptical region are masked with zeros. Finally, a fully connected layer is used to extract five key points: left and right eyes, nose tip, and left and right corners of the mouth.

2.2 Face Recognition

For the face recognition scheme, this paper maps the extracted face regions to the deep feature maps of the face detection network. These regions are then convolved again by convolution layers to obtain embedding feature vectors. The obtained vector is compared with the face embedding feature vectors recorded for residents in the community [9] to determine if the person is an internal member of the community. The process is illustrated in Figure 7 as follows:



[Fig. 7] Schematic diagram of face feature comparison

When comparing the embedding feature vectors, they are mapped to a geometric space, and the user's identity is matched based on the Euclidean distance in the space. If the calculated distances between the vectors and the face feature results in the database are all below a set threshold, the person is identified as an abnormal individual [10-11]. Considering both the user-entered facial information and the faces captured by surveillance cameras, there is a certain "facial distortion" issue due to different shooting angles. In order to better match the faces, this paper applies an affine transformation to the user-entered facial information to align it with the facial angles captured by the cameras.

3. Experimental and Data Analysis

In this section, we will introduce the dataset used in this study and the objective evaluation metrics employed. Additionally, we will conduct ablation experiments and cross-comparison experiments to validate the superiority of the network architecture and overall detection performance. During the experimental phase, to ensure the stability of the results, all network models were tested 10 times on the test set, and the average of the test results was taken as the final result.

3.1 Dataset and Evaluation Metrics Configuration

In this experiment, , as there is no publicly available dataset specifically relevant to the experimental scenario in this paper, the authors have communicated and negotiated with the Longyin Shuijun Community's property management and homeowners' committee. With their permission, the community's surveillance and access control facial databases can be used to create the dataset in the experimental environment. In the data set production environment a total of 2174 images containing frontal face images of pedestrians were extracted from the surveillance footage of the community. These images were used for training and testing the face detection network. Additionally, 100 user face data from the community access control system were utilized to train and test the face recognition network by matching the faces in the surveillance footage.

Next is the configuration of performance evaluation metrics. In terms of the face detection network, the following metrics were used in this study:

Precision: Evaluates the classification accuracy of the network for detecting positive and negative samples in the detection area. It is calculated using the formula:

$$Precision = \frac{TP}{TP + FP}$$

Recall: Evaluates the network's ability to detect targets comprehensively. It is calculated using the formula:

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: It is a metric that combines precision and recall to provide a comprehensive assessment of a network model's performance. The formula for calculating F1-Score is as follows:

$$F1 = \frac{2 \times \mathrm{P}recision \times Recall}{\mathrm{P}recision + Recall}$$

Average Precision (AP): It reflects the accuracy of the network in detecting targets. The formula for calculating AP is as follows:

$$AP = \int_0^1 P_{smooth}(r)dr$$

Parameter Count: It indicates the total number of parameters in the network. The formula for calculating parameter count is as follows:

$$Parameter = (C_{in} \times K^2 + 1)C_{out}$$

For the face recognition module, this paper uses the accuracy metric, which is calculated using the following formula:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

In the above formulas, TP, TN, FP, and FN represent the number of predicted results in the confusion matrix. C_{in} and C_{out} represent the dimensions of input and output convolutional kernels, respectively. K represents the size of the convolutional kernel.

3.2 Ablation Experiment

In the ablation experiment, the network structure and performance of the face detection network were analyzed by dissecting the network modules. For the baseline comparison network, this paper used ResNet50 combined with an anchor-based detector. In the face recognition module, the spatial Euclidean feature distance method used in this paper was uniformly adopted. In the ablation experiment, the following five networks were set up for comparative analysis:

Network 1: Utilizes ResNet50 as the feature extraction network and an anchor-based detector.

Network 2: Utilizes an encoder-decoder network built with residual convolutional modules as the feature extraction network and an anchor-based detector.

Network 3: Utilizes a standard Transformer module as the encoder and residual convolutional modules as the decoder, along with an anchor-based detector.

Network 4: Utilizes a Transformer-Lite module as the encoder and residual convolutional modules as the decoder, along with an anchor-based detector.

Network 5: Utilizes a Transformer-Lite module as the encoder and residual convolutional modules as the decoder, while replacing the detector with an ellipse region-based Anchor-Free mechanism.

According to the evaluation metrics defined in section 3.1, the results of testing the aforementioned networks are shown in Table 1 as follows:

〈Table 1〉 Comparative Test Results of Ablation Experiment

	Precision	Recall	F1-Score	AP	Parameter/Mb	Accuracy
Network1	92.24%	89.49%	90.85%	91.38%	35.38	88.36%
Network2	94.83%	91.19%	92.97%	93.50%	42.16	90.15%
Network3	97.12%	94.08%	95.58%	94.27%	85.42	91.78%
Network4	97.34%	94.52%	95.91%	96.39%	47.53	92.29%
Network5	98.57%	95.85%	97.19%	97.19%	49.48	95.84%

Decoding network, the target area is amplified. The experimental results prove that enlarging the feature map improves the precision and recall metrics of the detection network, and accordingly, the final face identification rate also increases. In Network 3, compared with Network 2, it adopts the Vision Transformer module in the encoder of the feature extraction network to enhance the network's ability to extract global features, searching for the target from a global

perspective, and then using the convolution module in the decoder for further refined feature extraction. Compared with the experimental data of Network 2, the recall rate has significantly increased, indicating that the network's ability to find the target has been strengthened. However, due to the computational mechanism of the Transformer module, the number of network parameters has also significantly increased. Network 4 is structurally identical to Network 3, but in Network 4, the Vision Transformer module is replaced with the Transformer-Lite module designed in this paper. From the comparison results of the two sets of networks, it can be seen that Network 4 has significantly reduced the number of parameters compared to Network 3, and is closer to Network 2, achieving some parameter simplification. Additionally, Network 4 has also seen slight increases in other metrics, which indicates that the introduction of feature map calculations in the self-attention computation of the Transformer-Lite module has enhanced feature association to a certain extent, thereby enhancing network performance. Finally, in Network 5, the detector module uses the Anchor Free elliptical region generation mechanism designed in this paper. The Anchor Free mechanism eliminates the maximum value calculation process of the original prediction frame, reducing redundant computations, while representing the detection frame as an ellipse to better fit the face area, reducing the interference of background factors in subsequent face identity recognition. According to the test results, the use of the Anchor Free mechanism firstly reduces the interference of judgment on the target's positivity and negativity due to the elimination of the generation of candidate boxes, which improves the precision metric, and the reduction of interference factors also greatly improves the accuracy of face identity recognition. In summary, the results of the ablation experiment respectively validate the improvements in the accuracy of face detection and recognition brought about by the network structure, the Transformer-Lite module, and the Anchor Free mechanism.However, it can be seen from the above table that the incorporation of the Transformer module has greatly improved the detection accuracy, but it has also increased the network complexity. The comparison between Network 5 and Network 3 illustrates that the Transformer-Lite module and the Anchor-Free mechanism designed in this paper can effectively reduce the parameters, but in subsequent usage, it relies on hardware devices with higher computational power to support model operations.

3.3 Horizontal comparative experiment

In this section, a horizontal comparative test will be conducted for face detection and recognition networks to validate the superiority of the network designed in this paper in the application of community surveillance face recognition. The networks referenced and their detailed comparative results are shown in Table 2 below.

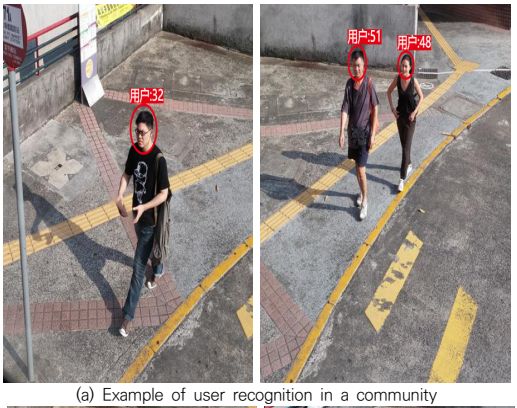
〈Table 2〉 Comparative Test Results

	Precision	Recall	F1-Score	AP	Accuracy
FaceNet[12]	91.59%	90.43%	91.00%	91.87%	90.12%
CosFace[13]	92.17%	92.59%	92.38%	93.44%	92.64%
ArcFace[14]	94.74%	95.62%	95.17%	95.67%	93.57%
DeepFace[15]	94.38%	94.28%	94.33%	95.35%	93.28%
our	98.57%	95.85%	97.19%	97.19%	95.84%

Based on Table 2, the network designed in this paper outperforms other networks in terms of F1-Score, AP, and Accuracy, which are comprehensive evaluation metrics. Compared to conventional face recognition tasks, the area of the face region in surveillance footage is smaller. However, no special processing is applied to the face extraction network in this paper. Instead, an encoding-decoding network is used to enlarge the face region, preventing feature loss.

Additionally, in the face recognition process, this paper aligns the face by extracting elliptical face regions and calculates the Euclidean distance in the feature space, reducing interference from background noise. As a result, the final face recognition accuracy metric, Accuracy, is also superior to other face recognition networks.

In conclusion, the feasibility of the network designed in this paper for user face identity recognition in small community surveillance footage has been verified through ablation experiments and cross-comparison experiments. The specific implementation results are shown in the following set of images.



(a) Example of user recognition in a community

(b) Sample of stranger user recognition

[Fig. 8] Examples of face detection and recognition.

Furthermore, the detection results from image (a) in the above set of images also verify that even in cases where there are differences in the angle between the user's face image and the captured face image, or when the face is partially occluded by wearing sunglasses, it is still capable of detecting and recognizing the facial identity information accurately.

4. Conclusion

This paper proposes a facial detection and recognition network for use in residential community security monitoring systems. It identifies individuals in surveillance footage to determine the presence of strangers entering and exiting the community, thereby enhancing the community's security level. To address the issue of small faces in monitoring scenarios, which are difficult to capture, a feature extraction network employing an encoder-decoder structure is designed. By upsampling to restore the size of feature maps in deep network structures, the loss of facial features is prevented.

Moreover, to enable the search for facial regions, a Vision Transformer module was incorporated. Based on this structure, a Transformer-Lite module was designed with fewer computation parameters and stronger feature area relatedness, enhancing the network's global feature extraction capability and quickly locating facial areas in surveillance footage.

Further, in the detection module, an anchor-free mechanism was used to generate an elliptical region to frame faces. Compared to the anchor-based schemes, the anchor-free mechanism reduces the generation of redundant boxes and lowers the false detection rate. Moreover, the elliptical region is more similar to the shape of a face. The extracted face region interferes less with other background features in the recognition stage. Finally, spatial Euclidean feature distance matching is used to determine facial identities. Through relevant experiments, the feasibility of the algorithm designed in this study has been verified for residential community facial recognition applications.

References

[1] W.Xin, L.Xiaonan, G.Huanbing, Z.Ziming, and Z.Yinlong, "Fall Detection Algorithm for Workshop Workers Based on Visible-to-Thermal Infrared Visual Monitoring", Control and Decision, pp.1-9, 2023.

[2] Z.Dexiang, W.Jun, Y.Peicheng, "Multi-scale All-scene Object Detection Method Based on Attention Mechanism ", Journal of Electronics and Information Technology, Vol.44, No.09, pp.3249-3257, 2022.

[3] X.Guishu, Z.Zihan, W.Yongchao, Z.Hongchao, and X.Weiji, "Transformer-based Face Recognition Method with Multi-level Feature Fusion ", Journal of Sichuan University (Natural Science Edition), pp.1-8, 2023.

[4] J.Ruirui, X.Yuhui, L.Fengkai, and M.Yuan. "Improved Visual Transformer-based Face Recognition Method ", Journal of Computer Engineering and Applications, Vol.59, No.08, pp.117-126, 2023.

[5] L.Jian, D.Jianiang, Z.Yanchen, and G. Yongkun. "A Survey on Transformer-based Object Detection Algorithms ", Journal of Computer Engineering and Applications, Vol.59, No.10, pp.48-64, 2023.

[6] L.Qingge, Y.Xiaogang, L.Ruitao, W.Siyu, X.Xueli, and Z.Tao. "A Survey on the Development of Transformer in Computer Vision ", Small and Micro Computer Systems, Vol.44, No.04, pp.850-861, 2023.

[7] L.Xiang, Z.Tao, Z.Zhe, W.Hongyang, and Q.Yurong. "A Survey on Transformer Research in Computer Vision ", Journal of Computer Engineering and Applications, Vol.59, No.01, pp.1-14, 2023.

[8] M.Yao, Zhimin, Y.Yanjun, and Pingping. "A Survey on CNN and Transformer Applications in Fine-Grained Image Recognition", Journal of Computer Engineering and Applications, Vol.58, No.19, pp.53-63, 2022.

[9] T.Shuang, H.Wei, and D.Yi. "Light-weight Face Recognition Method with Feature Masking ", Journal of Inner Mongolia University (Natural Science Edition), Vol.54, No.03, pp.272-280, 2023.

[10] L.Yunhong, L.Xingrui, X.Rongrong, S.Xueping, Z.Leitao, and B.Xiaohua. "Low-Resolution Face Recognition Based on Super-Resolution Reconstruction and Public Feature Subspace", Journal of Northwest University (Natural Science Edition),Vol.53, No.02, pp.241-247, 2023.

[11] Y.Jie, H.Jin, and S.Haode, "Masked Face Recognition Method Based on Lightweight Network", Electronic Measurement Technology, Vol.46, No.06, pp.159-165, 2023.

[12] F.Schroff, D.Kalenichenko, and J.Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering", IEEE, DOI:10.1109/CVPR.2015.7298682, 2015.

[13] H.Wang, Y.Wang, and A.Zhoul, "CosFace: Large Margin Cosine Loss for Deep Face Recognition", computer vision and pattern recognition, pp.5265-5274, 2018.

[14] J.Deng, J.Guo, and S.Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition", DOI:10.48550/arXiv.1801.07698, 2018.

[15] I.Perov, D.Gao, and N.Chervoniy, "DeepFaceLab: A simple, flexible and extensible face swapping framework", DOI:10.48550/arXiv.2005.05535, 2020.

담 하 의(He-Yi Tan) [정회원]

- September 1999 - July 2003, Hubei Minzu University, Bachelor's Degree in Computer Science and Technology.
- September 2006 - December 2008, Graduated with a Master's degree in Software Engineering from Sichuan University.
- March 2020 - Present, has been pursuing Ph.D. in Intelligent Fusion in IT at Mokwon University, Daejeon, Korea.

〈관심분야〉

Computer software development, application of artificial intelligence technology, robot control technology

민 병 원(Byung-Won Min) [정회원]

- He received M.S. degree in computer software from Chungang University, Seoul, Korea in 2005.
- He received Ph.D. degree in the dept. of Information and Communication Engineering, Mokwon University, Daejeon, Korea, in 2010.
- He is currently a professor of Mokwon University since 2010.

〈관심분야〉

digital communication systems, Big Data