International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 1, February, 2025, pp. 143-152
Journal homepage: http://iieta.org/journals/ts

Enhanced Face Identification Performance Using Online Mining Strategy in Multi-Task Cascaded Mask Convolutional Networks

Krishnaraj Mony*, Jeberson Retna Raj

Department of Computer Science and Engineering, School of Computing, Sathyabama Institute of Science and Technology, Chennai 600 119, Tamil Nadu, India

Corresponding Author Email: mailtokrish2023@gmail.com

https://doi.org/10.18280/ts.420113

Received: 7 May 2024
Revised: 19 September 2024
Accepted: 30 October 2024
Available online: 28 February 2025

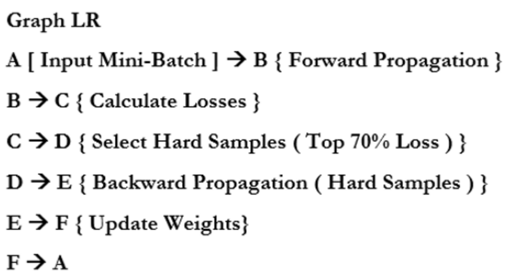
Keywords:
Age-Invariant Face Recognition (AIFR), cross-age image collection, Deep learning, face detection, hard sample mining, Multi-task Cascaded Mask Convolutional Network (MTCMCN), real-time performance, symmetry and occlusion handling

ABSTRACT

Due to the variety of lighting, postures, and occlusions, symmetry of faces and identification in an unrestricted area are difficult. The latest study demonstrates that deep learning techniques can do remarkably well on these two challenges. The complex transmitted multi-task structure the developers provide in this research takes advantage of the natural relationship between them to improve efficiency. The suggested Multi-task Cascaded Mask Convolutional Network (MTCMCN) has three layers of carefully planned deep convolution networks that work together to figure out where faces and landmarks are from a wide range of angles. Additionally, they provide a novel, continuous, difficult sample mining approach for learning procedures, which may automatically boost efficiency without the manual choice of samples. The use of a sizable cross-age image collection containing gender and age descriptors advances the creation of Age-Invariant Face Recognition (AIFR) and FAS. MTCMCN outperforms existing methods by achieving state-of-the-art accuracy on benchmarks like FDDB and WIDER FACE, exceeding 95% accuracy in some cases. It has a Central Processing Unit (CPU) speed of 16 frames per second and a GPU speed of 99 frames per second, ensuring real-time performance. The proposed system achieves this by using a special identification conditional block and live hard sample mining, thereby improving face recognition regardless of age.

1. INTRODUCTION

Numerous investigations, particularly in unregulated settings, have focused on automated facial expression recognition [1]. Expression research has several significant applications, including human-computer interaction, smart recording devices, and depressive and pain identification. This study describes a unique technique for recognizing facial expressions in both natural and lab-controlled photographs [2]. Here are a few conventional approaches for extracting visual features, which revolve around the use of handcrafted filters and the subsequent collection of mathematical data on patterns through histogram calculations. Arnable characteristics gradually replace the manually created ones, yet the histogram significantly influences the representation of statistical data in a small feature vector. To be able to recognize emotions [3] on someone's face when things aren't under control, you have to train a very complicated goal function that can handle big changes, like a person's head pose, as well as small changes in how they look [4]. Instance-based training provides a suitable framework to acquire complex functions illustrated by several simpler local approximations. Instead of immediately categorizing information, we employ a deep metric learning method to identify facial expressions, which teaches the network to compare the information. The model can use a new distance criterion to classify the vector of features produced by the neural network. In traditional facial expression recognition systems, every sample used for training consists of a facial image and the labeled expression that goes with it. The present research combines identical and different facial photos to create each training set. This eliminates the impact of an unbalanced sample size across classes. Additionally, this strategy mitigates the impact of inadequate training information, as the set of data contains a greater number of these combinations than there is training information.



```
Graph LR
A [ Input Mini-Batch ] ➔ B { Forward Propagation }
B ➔ C { Calculate Losses }
C ➔ D { Select Hard Samples ( Top 70% Loss ) }
D ➔ E { Backward Propagation ( Hard Samples ) }
E ➔ F { Update Weights}
F ➔ A
```

Figure 1. Flow of online hard sample mining for face recognition

For MTCMCN, the inclusion of live, difficult sample mining is a significant advancement because it greatly improves model performance and training. Conventional challenging sample mining techniques identify difficult instances before the training process starts, as they operate offline. Sometimes this entails hand-picking a collection of hardy samples, which takes a lot of time and effort. Furthermore, these pre-selected samples never vary during the training process, so they might not be able to adjust to evolving data ranges. Either way, MTCMCN employs an automated and dynamic technique called live difficult sample mining. This approach integrates difficult sample selection into the training loop, eliminating the need for manual labor and offline recognition. During every minibatch run, the network determines the loss function for each sample through forward propagation. The loss function, like the placement of face landmarks, measures how far off the expected output is from the true objective values. The "hard samples" are clearly the ones having the highest loss values right away. Typically, we select 70% of samples with the highest loss values. Then model and use these challenging samples, representing the harshest cases, for weight adjustments and backpropagation. Figure 1 indicates the representation of flow for online hard sample mining with face recognition. This continuous and adaptable approach has several advantages, including efficiency, objectivity, and a customizable nature. Emphasizing the most pertinent instances, online difficult sample mining improves the model's learning ability. This results in improved face recognition and alignment even in cases when the photographs are not particularly clear or the subjects are in odd stances or low-resolution images.

2. LITERATURE SURVEY

The deep neural networks have recently demonstrated higher capabilities in areas such as facial expression identification. Various studies [5] use Convolutional Neural Network (CNN) histogram analysis. This study sets itself apart from previous works by aiming to teach a histogram-based CNN using histogram-friendly chi-squared criteria [6]. The model modified a chi-squared separation by defining a learnable matrix as the fully connected layer [7]. The amended formula then trains the neural network, which generates an additional activation function using a histogram-based loss function [8].

A lack of adequate training information is one of the issues with deep learning for facial expression identification that leads to overfitting on training information [9]. Researchers have come up with two main ways to solve this problem: using 3D face modeling to improve the data and transfer machine learning to set up the convolutional parts of the proposed network [10]. Convolutional neural networks use the synthesized information as a technique to expand their dataset for learning. This study synthesizes new faces with different head positions and lighting, utilizing 3D modeling for each facial representation to instruct the neural network. The system creates the suggested neural network model based on histograms from 2D images [11]. Therefore, this method of incorporating 3D data generates fresh 2D facial images, potentially enhancing the effectiveness of the suggested neural network in managing these alterations when they aren't under control. The system may use previously trained deep neural networks for facial expression recognition. Although these networks offer a wealth of accessible variables, their development stemmed from distinct objectives [12]. Neural variables make it challenging to effectively train neural networks, as does the dearth of instructional content for facial emotion detection. This research [13] selects a large neural network with extensive face recognition datasets and uses it to train a smaller neural network on facial expression data. To increase the detector's capabilities while learning, rigorous sample mining is required.

On the other hand, traditionally difficult sample mining frequently takes place offline, which greatly increases the amount of physical work [14]. The model aims to develop an online computational sample mining approach for face alignment and recognition that can seamlessly integrate with the existing training method. In this research, the system proposes a novel framework that enables simultaneous learning of these two tasks using integrated cascaded CNNs [15]. The system has divided the suggested CNNs into three stages. In the first stage, a shallow CNN quickly generates candidate windows. Next, a more complex [16] refines the windows to eliminate a significant proportion of non-faced windows. Utilizing a stronger CNN, it then refines the findings and outputs the locations of the facial landmarks. This multi-tasking learning approach allows for a significant improvement in algorithmic efficiency. In this field of study, the MTCNN method is one of the most popular FD techniques. This method is a more detailed version of the CNN cascade-based FD algorithm, which works at different decision points, quickly gets rid of background noise in low-resolution stages, and carefully chooses candidates for the high-quality stage in the last step [17]. R-Net, P-Net, and O-Net are the three different phases of the MTCNN approach for performing face identification and landmark placement [18]. As per SOTA face database WIDER FACE, this method's FD reliability is 85.10% [19]. The SCRFD FD method discovered two essential components of ideal FD systems like the random sampling of training information and the distribution of computation. Based on these conclusions, the model proposes two strategies [20]. One method, known as "Sample Redistribution," expands the initial training data for the crucial stages using data from the conventional database [21].

A different method, known as computational redistribution, redistributes three key FD model parameters—the head, backbone, and neck—for calculation on the basis of a precise search approach [22]. This method claims to have the highest FD reliability to date, with a score of 96.06% on the WIDER FACE dataset. Formation and evaluation are the two key phases in traditional FR systems. In the learning step, the system first processes the input information before using methods for feature extraction to identify facial traits. The system then uses an attribute theme to store the characteristics [23]. The testing phase preprocesses the input facial information and obtains characteristics in a manner similar to that of the training phase. Researchers compare the extracted features from the testing information with the previously stored characteristics in the characteristic templates [24]. Early studies on face recognition primarily focused on algorithms that compared fundamental aspects of facial feature architecture through photographic processing methods [25]. By employing specialized contour detectors and borders, the researchers attempted to locate face features and assess their distances and locations between them [26]. In FR methods, researchers thoroughly investigated the possibility of using facial landmarks and their arrangement. Once 3D landmarks provide depth data, geometry-based techniques become more effective in 3D FR methods.

In subsequent studies using what the model refers to as holistic FR approaches, experts used the entire facial region as input for FR methods. Instead, the researchers employed characteristics that could describe the image textures at different points as a result of advancements in methods for computer vision [27]. This feature-based method matches the local characteristics of photos for FR purposes. Researchers refer to the combined or hybrid method for FR structures, which combines feature-based and holistic strategies to increase efficiency [28]. Up until the advent of DL-based approaches, these were the usual techniques primarily employed in FR devices. Over the past several years, the application of deep learning and CNN for facial recognition techniques has had significant effects on performance improvement [29]. With the development of highly sophisticated architectures and discriminating approaches to learning, the performance of FR methods had reached an astonishing level. DL approaches, when trained on significant volumes of data, can create FR systems that are resilient to various training information [30]. Convolution layers, data pooling layers, and completely connected layers make up the majority of a CNN structure. A convolutional layer attempts to identify facial traits based on the provided information.

The convolutional layer employs nonlinear transmission operations and a filter kernel to carry out the convolution process [31] by combining the outcomes for one layer's neuron groups into the specific neuron in the following layer, the pooling layers' goal is to reduce the size of the map of features. Implementing this CNN-based representation of feature algorithms in FR systems has significantly increased their effectiveness. The DL-based techniques also incorporate the two steps found in conventional [32]. During the training stage, we process the input information beforehand to adjust the input tensor for the neural network architecture, generating a unified include map. This includes adjusting the total number of images, height, width, and various other variables. The model may also resize or change the position of the input data [33].

New deep learning methods, especially the utilization of Multi-Task Cascaded Mask Convolutional Networks (MTCMCN), have recently enabled significant advances in face recognition. The author [34] claim that these networks help one to recognize persons while also handling other issues such as occlusions and light fluctuations. MTCMCN has made significant strides with its online mining approach. This approach, according to researcher [35], uses harsh sample mining to enhance learning and thereby helps the model to function without human intervention. Research indicated up to 2024 that these techniques not only improve face recognition accuracy but also AIFR performance [36]. These methods have improved by using entire cross-age databases listing both gender and age, ensuring reliable findings on common datasets such as FDDB and AFLW [37]. These developments have led to a significant demand for more accurate and flexible identification systems. As a result, MTCMCN is a leading framework in current face identification technologies.

3. PROPOSED METHODOLOGY

Figure 2 shows our method's complete pipeline. Researchers first resize an image to different scales to generate an image pyramid, which serves as the input for the following three-phase cascaded framework:

- To get the prospective windows and associated boundary-box regression vectors, the model implements an entirely convolutional network dubbed the suggestion network. The proposed system calibrates the choices using the calculated regression bounding box matrices. After that, the model utilizes non-maximum suppression to combine options with significant overlaps.
- The system sends each applicant to the Refine Network, a separate CNN that calibrates using bounding box regression, combines NMS candidates, and gets rid of even more wrong members.
- This stage focuses on providing a more detailed description of the face, similar to the subsequent stage. The network will specifically output the locations of five face features.

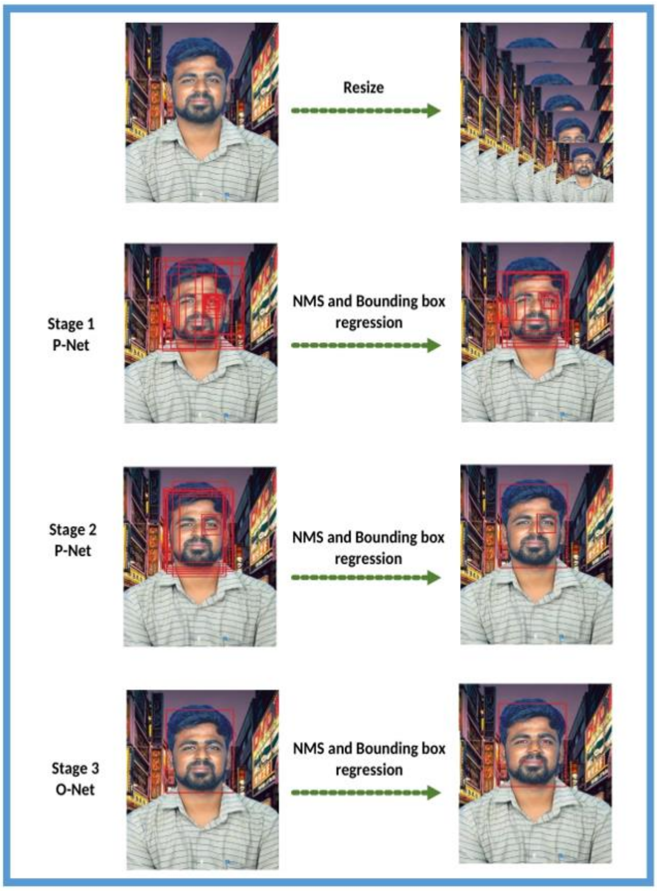

Resize
Stage 1
P-Net
NMS and Bounding box regression
Stage 2
P-Net
NMS and Bounding box regression
Stage 3
O-Net
NMS and Bounding box regression

Figure 2. The proposed MTCMCN framework's pipeline

3.1 Proposed MTCMCN architecture

In several CNNs, face recognition has been implemented. However, we observed that the following details might restrict its performance: Certain filters might be unable to provide discriminatory descriptions because their weights are not diverse enough. Identifying faces is a difficult binary classification problem; therefore, it may require fewer filters but a greater number of them than other multi-class objection recognition and categorization tasks. To do this, we cut down on the number of filters and switch from a 55 filter to a 33 filter for simpler computing while boosting efficiency. Compared to the prior construction, these advancements allow for improved performance with shorter runtimes, as demonstrated in Table 1.

Table 1. Comparison of speed and validation accuracy of the proposed MTCMCN and existing system

Group	MTCMCN	300 Times Forward	Accuracy (%)
Group1	12-Net [19]	0.041 sec	95.1
Group1	P-Net	0.033 sec	94.8
Group2	24-Net [19]	0.745 sec	95.3
Group2	R-Net	0.512 sec	95.6
Group3	48-Net [19]	3.602 sec	94.2
Group3	O-Net	1.352 sec	95.6

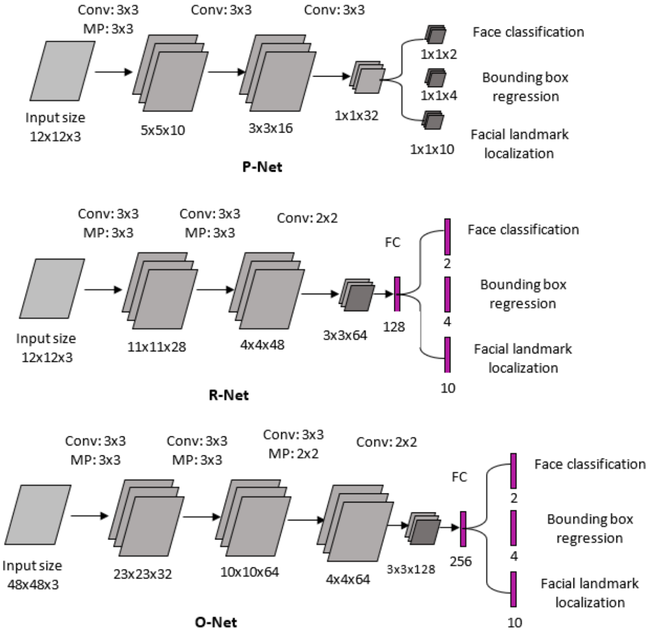


Figure 3. MTCMCN architecture

Figure 3 indicates the complete CNN architecture for the proposed system. The MTCMCN's cascaded design facilitates learning hierarchical characteristics and increases efficiency. Multi-task learning produces new knowledge and shared representations. Several drawbacks, including overfitting and the expense of computational capability, might be considered. But by concentrating on difficult instances, online hard sample mining increases dependability. More theoretical analysis—including ablation studies and links to significant frameworks—would strengthen the assertion made in the study about MTCMCN working. Researchers use 3 tasks—non-face/face categorization, bounding box regression analysis, localization of facial landmarks—to train our MTCMCN detectors. The goal of the lesson is presented as a two-class categorization issue. We employ the cross-entropy loss for every sample p_x:

$$L_x^{det} = -\left(j_x^{det}\log(p_x) + (1 - j_x^{det})(1 - \log(p_x))\right) \tag{1}$$

where, *x* is the chance that an example is a face, as determined through the network. The ground-truth label is indicated by the notation j_x^{det}. We forecast the offset between every potential window and the closest reality for every window. Everyone uses the Euclidean loss for every sample Xi and frame to achieve the learning objective as the following issue:

$$L_x^{box} = |\hat{j}_x^{box} - j_x^{box}|_2^2 \tag{2}$$

where, $\hat{j}_x^{box}$ is the ground-truth coordinates and j_x^{box} is the regression goal that was derived from the network. There are 4 coordinates—left top, width, height, &$\hat{j}_x^{box}$—and as a result, j_x^{box}∈R.

Facial landmark identification is formulated as a regression issue, much like the bounding box regress assignment, and we want to minimize the Euclidean loss:

$$L_x^{LM} = |\hat{j}_x^{LM} - j_x^{LM}|_2^2 \tag{3}$$

where, $\hat{j}_x^{LM}$the ground truth is coordinate &j_x^{LM} is the network coordinator for the facial landmark. There are five facial landmarks: the right eye, the left eye, the left corner of the mouth, the nose, and the right corner of the mouth j_x^{LM}∈R.

Training management uses various types of training images, including non-face, face, and partially aligned face images, as each CNN performs a distinct function. This scenario does not utilize all the loss functions. For example, the developers simply compute the background region sample and leave the other two losses at 0. You can use an example-type indication to accomplish this immediate task. The overall learning goal can then be stated as

$$\min \sum_{x=1}^{N} \sum y \in \{det, box, LM\} \alpha_y \beta_x^y L_x^y \tag{4}$$

where, *N* is the total of training samples. a_j denotes the task's importance. Researcher use α_y in P-Net &L_x^y in R-Net, while in O-Net of more accurate facial landmarks localization. β_x^y is the sample type indicator. It makes sense in this situation to develop the CNNs using random gradient descent.

Once the initial classifiers have been trained to adapt to the learning procedure, the system performs online difficult sample extraction in the face categorization job, instead of traditional hard sample mining. Everyone specifically sorts the loss computed in the propagation forward stage from all samples from every mini-batch and chooses the top 70 percent of them as difficult samples. At that point, we just compute the gradient from the difficult samples in the backward transmission stage. This means that during training, we overlook the simple data that are less useful in improving the detector. The results of experiments demonstrate that this approach performs better without manual sample selection.

3.2 AIFR

Figure 4 illustrates the architecture of the proposed MTCMCN. The major issue with AIFR was that age variability typically introduces widening intra-class gaps because faces vary dramatically over time. Due to the strong entanglement of unrelated data, such as modifications to facial shape and appearance, a significant separation between the two faces of the same person makes it challenging to distinguish them. Formally, their linear factorization module: Given a feature vector i∈$\boldsymbol{i}_{id}$derived from an input image I∈ R 3XHXW.

$$i = i_{age} + i_{id} \tag{5}$$

where, i_{age} and i_{id} denote the age and identity-related components, respectively.

Instead, researchers suggest using attention-based feature decomposition, also known as AFD, to break down the mixed feature maps at the high-level semantics level to overcome these shortcomings. Aging/rejuvenation impacts, including beards and wrinkles, appear within the semantic component space but disappear in the one-dimensional characteristics; operating on feature vectors is more challenging than on feature mappings. Formally, researchers utilize a ResNet-like backbone as encoder σ to extract mixed feature maps X ∈ R C×H0×W0 from an input image I, i.e. X = $\sigma(I)$, the AFD could be defined as follows:

$$I = \left(I \circ \sigma(I) I_{age}\right) + \left(I \circ \left(1 - \sigma(I)\right) I_{age}\right) \quad (6)$$

The symbol I represents the attention modules and denotes the division of elements. In this process, an age estimate assignment controls the focusing module, which removes age-related data from the feature maps. A face identification challenge guides the remaining portion, believed to contain identity-related data.

As a result, the attention mechanism constrains the decomposition modules, enhancing their effectiveness in identifying age-related characteristics in semantic maps of features. Skipping interconnections between the decoder and the encoder maintains the remaining data, which is crucial for FAS. Note that the two relevant tasks only expect X to include the age and identification data.

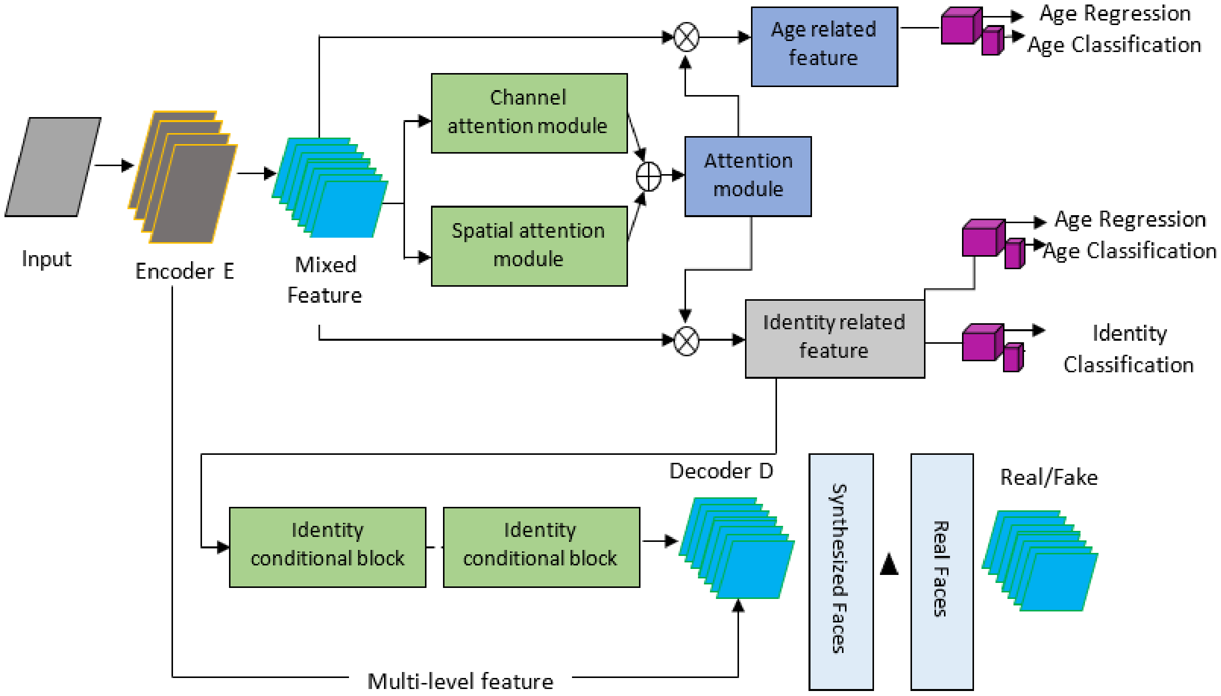


Figure 4. Architecture of proposed AIFR

3.3 Conditional module

The mainstream face-aging research typically divides the ages into many non-overlapping age groups due to the subtle changes in appearance over time with tiny age gaps. As shown in Figure 5(a), these techniques commonly utilize one-hot encoding to identify an age group of interest to regulate the rejuvenation/aging processes. The application of the one-hot age circumstance teaches every age category the group-level rejuvenation/aging pattern, including individuals who start growing beards at 30 years old. The model proposes the Identity Conditional Block (ICB) to achieve an identity-level rejuvenation/aging pattern, which addresses the issues caused by one-hot encoding. Additionally, it incorporates the weight-sharing technique to improve the smoothness of the ageing of synthesized faces. To learn an identity-level rejuvenation/aging pattern, the proposed ICB uses the identity-related characteristic from AFD as inputs. Then, as illustrated in Figure 5(b), we propose the weights-sharing technique, which enhances the age-synthesized faces' smoothness by sharing some convolution filters between adjacent age groups. The argument supports this theory, stating that faces gradually alter as we age, and that similar filtering can pinpoint shared aging and rejuvenation patterns among individuals of similar ages living in close proximity.

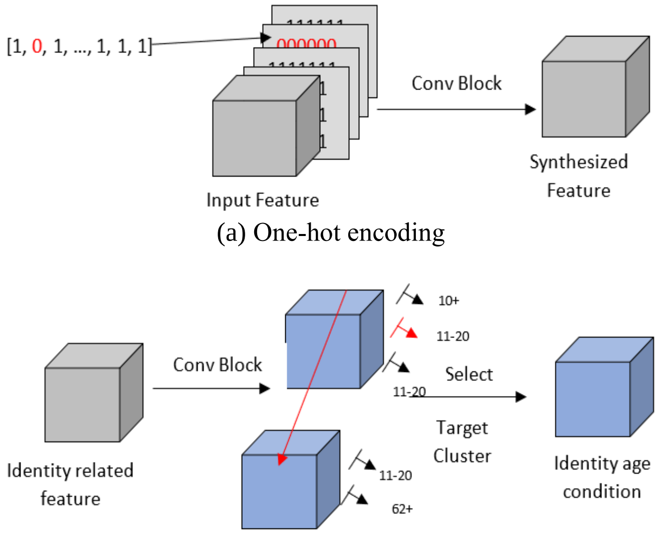
The following is a definition of the loss function to optimize age estimation:

$$\begin{aligned} l_{AE}(I_{age}) = E_X \Big[& l_{MSE}\left(DEX\left(A(I_{age})\right) j_{age}\right) \\ & + l_{CE}\left(A\left(W(I_{age})\right) C_{age}\right)\Big] \end{aligned} \quad (7)$$

AIFR's overall loss is expressed as follows: where I_{age} was the identification label, and λ ∗ regulates the balance of various loss phrases and first phrase was the Cos Face loss, the second term is the aged assessment loss, and the last phrase is the domain adaptation loss. Remaining two terms, which contain various inputs and have been taught separately, employ the same network topology.

$$\begin{aligned} L^{AIFR} = & \, l_{COSFACE}\left(\left(A(I_{age})\right), j_{age}\right) \\ & + \lambda^{AIFR}_{age} \, l_{AE}(I_{age}) \\ & + \lambda^{AIFR}_{id} \, l_{AE}\left(GRL(I_{age})\right) \end{aligned} \quad (8)$$

For simplicity, the batch normalization and activation functions were disregarded, and our face identification model was generated purely according to the settings (apart from the AFD). Applying the identity conditional module (ICM) with the string for ICBs allows for the precise derivation of the individual level age requirement from the prejudiced facial information Xi.



(a) One-hot encoding

(b) Identity conditional block

Figure 5. Comparison of ICB and one-hot encoding

3.4 Optimization and inference

Face can improve the model's comprehension of AIFR because it teaches discriminating facial descriptions and age estimates, whereas the FAS generates visual outputs. Therefore, by optimizing these two jobs in a manner akin to a GAN, we can jointly complete both tasks, as they mutually benefit from each other. In other words, FAS can assist in gathering identification-related characteristics and enhance the model interpretation ability of AIFR, while AIFR encourages FAS to render faces to safeguard its own identity. As a result, researchers train these two jobs alternatively using a single, multi-task, point-to-point structure.

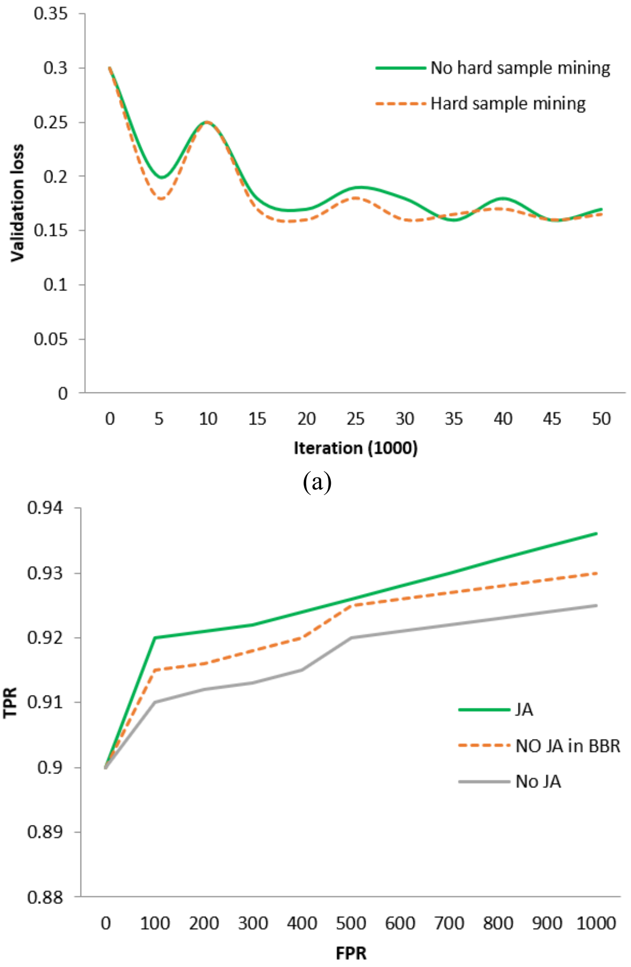
4. RESULTS AND OBSERVATIONS

First, the model measures how well the proposed difficult sample mining approach works. Then, using modern techniques from the face detection dataset and benchmark, wider faces, and marked facial landmarks in the wild benchmark, it evaluates the face detectors in addition to alignments. The experiments were conducted on two well-known standard datasets for face recognition, FDDB and WIDER FACE. The FDDB collection contains 5,171 labeled face images across 2,845 pictures. WIDER FACE is the harder dataset. The dataset consists of 32,203 shots and 393,703 labeled face-bounding boxes. We randomly split each dataset into three sets: training, validation, and testing. For each set, the splits were 80:10:10. The proposed approach used bilinear interpolation to resize the pictures to the same dimension of 224×224 pixels. We used random cropping, horizontal shifting, and color disturbance to make the model more resistant to changes in lighting and direction. The proposed system used the test set to improve the MTCMCN hyper parameters. The model found the Adam algorithm to be the best, with a learning rate of 0.001 and a batch size of 32. There was a total of 100 epochs of training for the model.

4.1 Training data

Here, researchers utilize four distinct kinds of information annotation in the learning procedure because recognizing faces and aligning are tasks that we execute together: Negative areas include those where any ground-truth face had an intersection-over-union ratio that was less than 0.3; positives include IoU above 0.65 to the ground-truth face, IoU between 0.4 and 0.65 to the part face, and faces labeled with locations of five landmarks. The system employs both negatives and positives for face tasks like classification, uses positives and partial faces for bounding box regression, and uses landmark faces for localizing facial landmarks. In three stages, the model describes the learning information for each network. The first step, P-Net, gathers positives, negatives, and partial faces by randomly cropping several patches from the WIDER FACE. After that, we cut out CelebA's landmark faces. The second part, known as R-Net, involves researchers recognizing landmark parts from Celebi and using our system's first step to identify faces from WIDER FACE, thereby collecting positive tests, negative results, and partial faces. O-Net, the final stage, gathers data in a manner akin to R-Net, but employs our structure's first two steps to identify faces.

4.2 Online hard sample mining



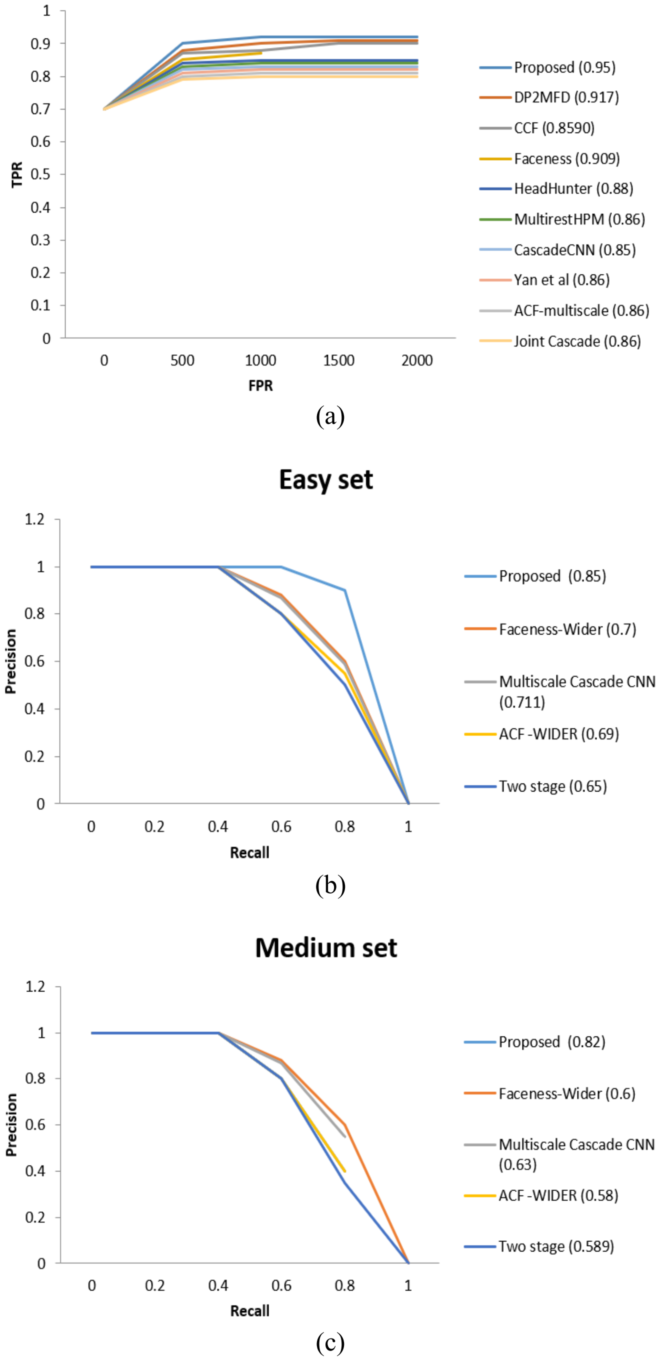
(a)

(b)

Figure 6. (a) O-Net validation losses with and without heavy sampling extraction (b) "JA" stands to represent joint face alignment

The model trains two O-Nets and contrasts their loss curves to assess the impact of the proposed online hard sample mining technique. To make comparisons more accurate, researchers exclusively train the O-Nets for the face categorization task. These two O-Nets have identical learning variables, except for the network startup. To make comparisons simpler, everyone employs an established rate of learning. Figure 6(a) displays the loss curves from two separate training techniques. Hard sample extraction is advantageous for boosting efficiency. Researchers compare the FDDB performance of two distinct O-Nets to assess the impact of joint identification and aligning. In these two O-Nets, the framework also contrasts the results of the bounding box regression method. According to Figure 6(b), joint landmarks in localization assignment training are advantageous for both face classification and bounding box regression assignments.

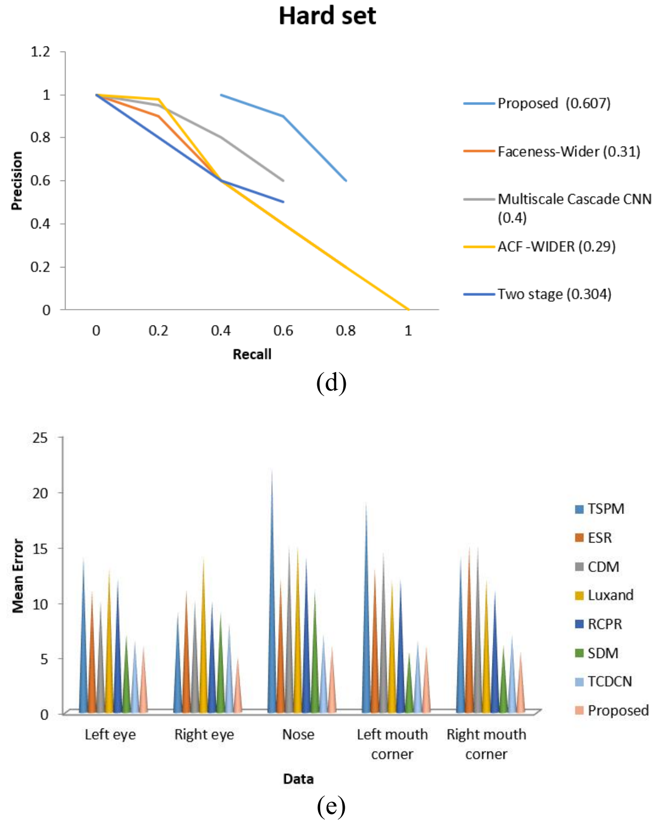
4.3 Evaluation



(a)

(b)

(c)



(d)

(e)

Figure 7. A FDDB assessment

The model compares it to both the modern techniques in FDDB and the cutting-edge techniques in wider face to evaluate how well our algorithm performs in detecting faces. Our strategy consistently beats all previous methods by a significant margin in both benchmarks, as shown by Figure 7 (a-d). The system also tests our strategy on a few challenging images.

Figure 7(a) presents an assessment of FDDB, while Figure 7(b-d) evaluates three subsets of WIDER FACE. The next number represents the technique's average precision. (e) An AFLW assessment for facial alignment.

4.4 Runtime efficiency

The proposed approach can detect and align joint faces very quickly because of the cascade architecture. It requires 16 frames per second on a 2.60GHz CPU and 99 frames per second on a GPU. The current implementation relies on un-optimized MATLAB code.

4.5 Evaluation on AIFR

For fair assessments, researchers assess AgeDB performance using the models developed on SCAF. Table 2 compares the verification reliability of the models we developed to existing state-of-the-art AIFR methods, showcasing the proposed technique's higher accuracy.

Similar to the LFW, the model applies the same technique, with 600 positive and negative pairs in each fold. The system uses LCAF to train the framework on this set of data, and Table 3 displays the results. Specifically, our approach outperforms the most advanced AIFR approaches currently in use, setting a new standard on the CALFW dataset. The cross-age celebrity dataset, which functions as a public age dataset for AIFR, encompasses 163446 expression photos of 2000 celebrities, captured in a remote setting with varying lighting, position, age, and other characteristics. Data collection by search engines clutters CACD with duplicated and incorrectly labeled photos. The proposed approach creates a thoroughly documented version known as CACD-VS, or CACD validation subset, to enable fair assessments. This version also adheres to Table 4, which compares the proposed approach to further modernization in CADCD-VS.

Table 2. Test outcomes on AgeDB-30

Method	Accuracy (%)
[10]	55.3
[12]	89.93
[13]	94.12
[14]	91.72
[15]	95.6
[16]	95.20
[17]	95.35
MTCMCN	96.44

Table 3. Test outcomes on CALFW

Method	Accuracy (%)
[18]	86.7
[19]	85.3
[21]	84.44
[22]	97.2
MTCMCN	95.66

Table 4. Test outcomes on CACD-VS

Method	Accuracy (%)
[10]	85.11
[12]	87.72
[13]	96.12
[14]	97.52
[15]	98.61
[16]	99.25
[17]	99.40
MTCMCN	99.58

Table 5. Test outcomes on FG_NET (Leave one out)

Method	Accuracy (%)
[10]	37.51
[12]	47.62
[13]	70.00
[14]	76.22
[15]	86.55
[16]	88.22
[17]	93.21
[21]	94.48
MTCMCN	94.83

Table 6. Test outcomes on FG-NET (MF1)

Method	Accuracy (%)
[36] FUDAN-CS-SDS	25.58
[26] SphereFace	47.61
[33] TNVP	47.75
[23] OE CNN	52.78
[22] DALL	57.99
MTCMCN	57.28

The proposed MTCMCN significantly outperforms the further advanced approach, resulting in a 0.15 increase over the most recent one. Table 5 shows the rank-one recognition rate. The research methodology significantly outperforms previous research. On the other hand, the MF1 encompasses roughly 1 million images from 690,000 distinct individuals, serving as distractions in the image set. The big and medium learning methodologies. Less than 0.5 million photos are required for the tiny protocol's set of training images. The model correctly follows the minimal method for assessing our newly developed model on FG-NET. Table 6 reports the experimental findings. Due to the enormous number of incorrectly labeled probes and collection face images in the MF1 distractors, our approach outperforms competing methods.

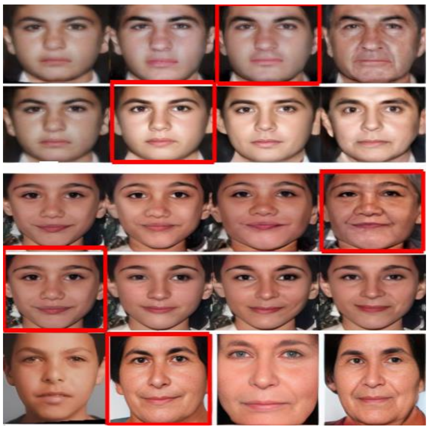


Figure 8. Qualitative results with proposed MTCMCN

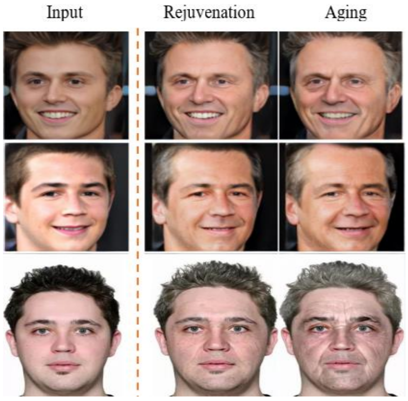


Figure 9. Qualitative evaluations of previous FGNET research

Figure 8 displays sample outcomes from the outside datasets LCAF, MORPH, and FG-NET. With high visual realism, our approach can imitate a face-age synthesis process among age groups. Despite changes in gender, expression, race, and occlusion, the synthesized faces remained photorealistic, retaining natural information in the muscle tissue, skin, and wrinkles, and reliably maintaining personal personalities. Figure 9 demonstrates the generalizability of the proposed approach. Researchers compare our findings qualitatively to earlier research, including CAAE and AIM on MORPH and FGNET. While our MTCMCN uses the same age condition to synthesize faces based on the multi-level characteristics derived from the encoder, AIM and CAAE both produce over smoothed faces as a result of their image reconstruction, as shown in Figure 8. Note that we directly cite rivals' results from their papers to ensure a fair comparison. The FAS literature extensively employs this practice to prevent bias or inaccuracy resulting from self-implementation.

Check out the appendix for more information on identity-conditioned module ablation research and how it compares to CAAE and IPCGAN in terms of two evaluation standards: age accuracy and identity protection.

5. CONCLUSION

In this research, researchers present a framework for joint face detection and alignment based on MTCMCN. Trial outcomes show that, while maintaining real-time efficiency, our methods regularly outperform the latest techniques across a range of difficult benchmarks. To further boost efficiency, we will be making use of the natural link between face detection and other face analysis activities. The model uses ICM for identity-level face age synthesis and AFD for segmenting the features into identity-related and age-related characteristics. Extensive facial recognition trials on benchmark and cross-age data sets demonstrate the effectiveness of our suggested approach. The future stage of study will concentrate on integrating multimodal data to increase robustness, refining feature segmentation, boosting flexibility via adaptive learning, and improving real-time processing on peripheral devices. The main future work is to investigate how 3D facial analysis can improve alignment and depth precision, as well as inclusion and equity across many groups. The system focuses on increasing accuracy, recall, and F1 scores through advanced algorithms and datasets. Efforts will include boosting adaptability, fairness across diverse groups, and exploring 3D face analysis for better alignment accuracy.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the Department of Computer Science and Engineering, School of Computing, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, for their invaluable support and resources that greatly contributed to the success of this research.

REFERENCES

[1] Mokalla, S.R., Bourlai, T. (2023). Utilizing alignment loss to advance eye center detection and face recognition in the LWIR band. IEEE Transactions on Biometrics, Behavior, and Identity Science, 5(2): 255-265. https://doi.org/10.1109/TBIOM.2023.3251738

[2] Feng, S.M., Nong, X.Z., Hu, H.F. (2022). Cascaded structure-learning network with using adversarial training for robust facial landmark detection. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18(2): 1-20. https://doi.org/10.1145/3474595

[3] Li, M., Huang, B., Tian, G. (2022). A comprehensive survey on 3D face recognition methods. Engineering Applications of Artificial Intelligence, 110: 104669. https://doi.org/10.1016/j.engappai.2022.104669

[4] Khalifa, A., Abdelrahman, A.A., Strazdas, D., Hintz, J., Hempel, T., Al-Hamadi, A. (2022). Face recognition and tracking framework for human—Robot interaction. Applied Sciences, 12(11): 5568. https://doi.org/10.3390/app12115568

[5] Wu, H. (2023). 3D Face feature processing and recognition technology. International Journal of Advanced Network, Monitoring and Controls, 7(4): 9-20. https://doi.org/10.2478/ijanmc-2022-0032

[6] Jayanthi, E., Ramesh, T., Kharat, R.S., Veeramanickam, M.R.M., Bharathiraja, N., Venkatesan, R., Marappan, R. (2023). Cybersecurity enhancement to detect credit card frauds in health care using new machine learning strategies. Soft Computing, 27(11): 7555-7565. https://doi.org/10.1007/s00500-023-07954-y

[7] Anghelone, D., Chen, C., Ross, A., Dantcheva, A. (2025). Beyond the visible: A survey on cross-spectral face recognition. Neurocomputing, 611: 128626. https://doi.org/10.1016/j.neucom.2024.128626

[8] Zhou, C., Zhi, R. (2022). Learning deep representation for action unit detection with auxiliary facial attributes. International Journal of Machine Learning and Cybernetics, 13(2): 407-419. https://doi.org/10.1007/s13042-021-01413-6

[9] Fegade, V., Chodankar, A., Bhingle, A., Mhatre, S. (2022). Residential security system based on facial recognition. In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, pp. 01-09. https://doi.org/10.1109/ICOEI53556.2022.9776940

[10] Jia, H.J., Xiao, Z.J., Ji, P. (2022). Real-time fatigue driving detection system based on multi-module fusion. Computers & Graphics, 108: 22-33. https://doi.org/10.1016/j.cag.2022.09.001

[11] Pradeepa, K., Bharathiraja, N., Meenakshi, D., Hariharan, S., Kathiravan, M., Kumar, V. (2022). Artificial Neural networks in healthcare for augmented reality. 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP), Bengaluru, India, pp. 1-5. https://doi.org/10.1109/CCIP57447.2022.10058670

[12] Vu, H.N., Nguyen, M.H., Pham, C. (2022). Masked face recognition with convolutional neural networks and local binary patterns. Applied Intelligence, 52(5): 5497-5512. https://doi.org/10.1007/s10489-021-02728-1

[13] Adeline, U.M., Gaspard, H., Innocent, K. (2022). A real-time face recognition attendance using machine learning. In International Conference on Applied Machine Learning and Data Analytics, Lübeck, Germany, pp. 91-107. https://doi.org/10.1007/978-3-031-34222-6_8

[14] Sheriff, M., Jalaja, S., Dinesh, K.T., Pavithra, J., Puja, Y., Sowmiya, M. (2023). Face emotion recognition using histogram of oriented gradient (hog) feature extraction and neural networks. Recent Trends in Computational Intelligence and Its Application, CRC Press, United States, pp. 114-122.

[15] So, J., Han, Y. (2023). Heatmap-guided selective feature attention for robust cascaded face alignment. Sensors, 23(10): 4731. https://doi.org/10.3390/s23104731

[16] Yan, C., Meng, L., Li, L., Zhang, J., Wang, Z., Yin, J., Zhang, J.Y., Sun, Y.Q., Zheng, B.L. (2022). Age-invariant face recognition by multi-feature fusionand decomposition with self-attention. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18(1s): 1-18. https://doi.org/10.1145/3472810

[17] Sadeghi, H., Raie, A.A. (2022). HistNet: Histogram-based convolutional neural network with Chi-squared

deep metric learning for facial expression recognition. Information Sciences, 608: 472-488. https://doi.org/10.1016/j.ins.2022.06.092

[18] Dosso, Y.S., Kyrollos, D., Greenwood, K.J., Harrold, J., Green, J.R. (2022). NICUface: Robust neonatal face detection in complex NICU scenes. IEEE Access, 10: 62893-62909. https://doi.org/10.1109/ACCESS.2022.3181167

[19] Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y. (2022). Probabilistic attribute tree structured convolutional neural networks for facial expression recognition in the wild. IEEE Transactions on Affective Computing, 14(3): 1927-1941. https://doi.org/10.1109/TAFFC.2022.3156920

[20] Oroceo, P.P., Kim, J.I., Caliwag, E.M.F., Kim, S.H., Lim, W. (2022). Optimizing face recognition inference with a collaborative edge–Cloud network. Sensors, 22(21): 8371. https://doi.org/10.3390/s22218371

[21] Vinod, D., Bharathiraja, N., Anand, M., Antonidoss, A. (2021). An improved security assurance model for collaborating small material business processes. Materials Today: Proceedings, 46: 4077-4081. https://doi.org/10.1016/j.matpr.2021.02.611

[22] Yu, M., Ju, S., Zhang, J., Li, S., Lei, J., Li, X. (2022). Patch-DFD: Patch-based end-to-end DeepFake discriminator. Neurocomputing, 501: 583-595. https://doi.org/10.1016/j.neucom.2022.06.013

[23] Ulukaya, S., Sandıkçı, E.N., Eroğlu Erdem, Ç. (2023). Consensus and stacking based fusion and survey of facial feature point detectors. Journal of Ambient Intelligence and Humanized Computing, 14(8): 9947-9957. https://doi.org/10.1007/s12652-021-03662-3

[24] Nagu, B., Arjunan, T., Bangare, M.L., Karuppaiah, P., Kaur, G., Bhatt, M.W. (2023). Ultra-low latency communication technology for Augmented Reality application in mobile periphery computing. Paladyn, Journal of Behavioral Robotics, 14(1): 20220112. https://doi.org/10.1515/pjbr-2022-0112

[25] Bulatovich, B.K., Vladimirovna, E.E., Vyacheslavovich, T.D., Sergeevna, S.N., Sergeevna, C.Y., Ming, Z.H., Burie, J.C., Muzzamil, L.M. (2022). MIDV-2020: A comprehensive benchmark dataset for identity document analysis. Компьютерная оптика, 46(2): 252-270.

[26] Mohammed, O.A., Al-Tuwaijari, J.M. (2022). Analysis of challenges and methods for face detection systems: A survey. International Journal of Nonlinear Analysis and Applications, 13(1): 3997-4015. https://doi.org/10.22075/ijnaa.2022.6222

[27] Hoo, S.C., Ibrahim, H., Suandi, S.A., Ng, T.F. (2023). Lcam: Low-complexity attention module for lightweight face recognition networks. Mathematics, 11(7): 1694. https://doi.org/10.3390/math11071694

[28] Roozbahani, K.M., Zadeh, H.S. (2022). Face detection from blurred images based on convolutional neural networks. In 2022 International Conference on Machine Vision and Image Processing (MVIP) Ahvaz, Islamic Republic of Iran, pp. 1-10. https://doi.org/10.1109/MVIP53647.2022.9738783

[29] Hossain, M.A., Assiri, B. (2022). Facial expression recognition based on active region of interest using deep learning and parallelism. PeerJ Computer Science, 8: e894. https://doi.org/10.7717/peerj-cs.894

[30] Abbaspoor, N., Hassanpour, H. (2022). Face recognition in a large dataset using a hierarchical classifier. Multimedia Tools and Applications, 81(12): 16477-16495. https://doi.org/10.1007/s11042-022-12382-5

[31] Dash, P., Kisku, D.R., Gupta, P., Sing, J.K. (2022). Fast face detection using a unified architecture for unconstrained and infrared face images. Cognitive Systems Research, 74: 18-38. https://doi.org/10.1016/j.cogsys.2022.03.001

[32] Gao, H.X., Wu, M., Chen, Z.H., Li, Y.W., Wang, X.Y., An, S., Li, J.Q., Liu, C.Y. (2023). SSA-ICL: Multi-domain adaptive attention with intra-dataset continual learning for Facial expression recognition. Neural Networks, 158: 228-238. https://doi.org/10.1016/j.neunet.2022.11.025

[33] Anghelone, D., Lannes, S., Strizhkova, V., Faure, P., Chen, C., Dantcheva, A. (2022). Tfld: Thermal face and landmark detection for unconstrained cross-spectral face recognition. In 2022 IEEE International Joint Conference on Biometrics (IJCB), Abu Dhabi, United Arab Emirates, pp. 1-9. https://doi.org/10.1109/IJCB54206.2022.10007992

[34] Rusia, M.K., Singh, D.K. (2023). A comprehensive survey on techniques to handle face identity threats: Challenges and opportunities. Multimedia Tools and Applications, 82(2): 1669-1748. https://doi.org/10.1007/s11042-022-13248-6

[35] Xiang, P., Wu, K., Lin, C., Bai, O. (2024). MTCAE-DFER: Multi-Task Cascaded Autoencoder for Dynamic Facial Expression Recognition. arXiv preprint arXiv:2412.18988. https://doi.org/10.48550/arXiv.2412.18988

[36] Meena, G., Mohbey, K.K., Indian, A., Khan, M.Z., Kumar, S. (2024). Identifying emotions from facial expressions using a deep convolutional neural network-based approach. Multimedia Tools and Applications, 83(6): 15711-15732. https://doi.org/10.1007/s11042-023-16174-3

[37] Truong, T.D., Duong, C.N., Quach, K.G., Le, N., Bui, T.D., Luu, K. (2023). LIAAD: Lightweight attentive angular distillation for large-scale age-invariant face recognition. Neurocomputing, 543: 126198. https://doi.org/10.1016/j.neucom.2023.03.059