การประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 48
The 48th Electrical Engineering Conference (EECON-48)
วันที่ 19-21 พฤศจิกายน 2568 ณ โรงแรมฟูราม่า จังหวัดเชียงใหม่

Volume Deep Face: A 3D Face Descriptor for Face Authentication System

Suttipat Srisuk[1*], **Damrongsak Arunyagool**[1*], **Kitchanut Ruamboon**[1], and **Pantre Kompitaya**[2]

[1]Faculty of Engineering, Nakhon Phanom University
[1]srisuk.s@gmail.com, damrongsaka@npu.ac.th, kitchanut.r@npu.ac.th
[2]Thatphanom College, Nakhon Phanom University, pantre@npu.ac.th

***Abstract*—In this paper, we introduce the Volume Deep Face (VDF), a novel face representation proposed for the face authentication system. VDF provides a fast and compact representation of faces using deep learning, enabling one to encode more distinctive features. Using our proposed method, images can be generated to form a 3D VDF representation or a 2D face descriptor (2DFD). The 3D VDF is created from multiple images in the training set, while the 2DFD is generated from a single image during the testing phase. The matching confidence is evaluated using our new volume matching. Our face authentication system is verified with extensive experiments on the XM2VTS database.**

I. INTRODUCTION

Image representation plays a crucial role in the face authentication system [1]–[5], [8]. They provide rich information in which the transformed space is more distinctive than the image space. Due to the similar structure of face images, it is very hard to recognize the faces directly on image space. In addition, the gray level of human face images normally changes from time to time, resulting in a degradation of the performance of the face authentication system. Eigenfaces and Fisherfaces are among the most common methods for face representation where dimensionality of data is reduced while minimizing the within-class variance [1], [2]. Local binary pattern (LBP) was also a popular method proposed for face representation with the ability to enhance the gray level of face images. LBP measures the difference between the central pixel and its neighbors, and then encodes the changes as a binary number. LBP enhances robustness against illumination changes. Eigenfaces, Fisherfaces, and LBP can be categorized as hand-crafted feature extraction methods. The core strength of the convolutional neural network lies in its ability to extract useful features using the convolution operator [7], [8], [10]–[13]. CNN provides a dual mode that works both as a classifier and as a feature extractor. In this paper, we propose a novel face representation, termed volume deep face (VDF) which is constructed from multiple 2D face descriptors. We also propose a robust method for face and eye detections. The original contributions of this paper are: 1) It generates the face representation in both 2D and 3D forms, which provides the capability of constructing distinctive features. 2) It proposes robust 3D volume matching for face authentication.

*Corresponding Author

II. THE PROPOSED FACE AUTHENTICATION SYSTEM

A. Overview of the System

An overview of our proposed face authentication system is shown in Fig. 1. In the training phase, we first detect the face and eye using two-stage face and eye detections, 2SYOLO. The eye positions are then used to align the face in which the degree of rotation of the two eyes is zero. Face is then cropped to the normalized size 130×200. The aligned and cropped faces are then fed to the CNNs where VDF is generated. In the test phase, a similar process of 2SYOLO is performed to obtain an aligned and cropped face. 2DFD is created from a single image for testing. The volume matching is used to measure the similarity between VDF and 2DFD, resulting in true or false identity.

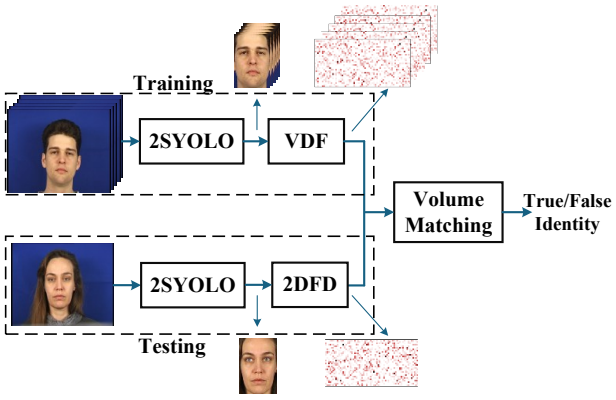

Fig. 1: The proposed face authentication system.

B. Face Detection and Alignment

The performance of a face authentication system can be significantly improved by precisely locating and aligning facial components. In particular, aligning human eyes across all face images ensures that key facial components are consistently positioned, enabling more reliable and accurate face verification. In this section, we introduce 2SYOLO, a two-stage YOLO-based approach for face and eye detection. In the first stage, the face bounding

boxes are located using YOLO-Face. In the second stage, eyes are localized within each detected face bounding box using YOLO-Eyes. Based on the detected eye positions, the face is then horizontally aligned to ensure that facial components are consistently positioned across all entire dataset. YOLO-Face is trained on face images, while YOLO-Eyes is trained specifically on eye region images to accurately detect eye positions within the detected face bounding boxes. Our two-stage YOLO-based approach is illustrated in Fig. 2. Our proposed

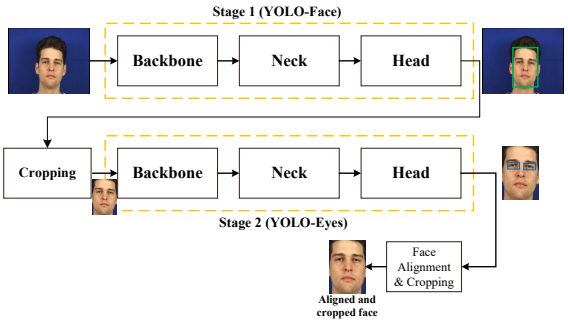

Fig. 2: Architecture of our two-stage YOLO-based approach for face and eye detection.

architecture is based on YOLOv6 [9] which has the following parts: a backbone, a neck and a head. The backbone is responsible for feature extraction and handles most of the network computation. The neck aggregates features from various stages of the backbone, enhancing the network's ability to capture multi-scale information, i.e. detecting objects at multiple scale. Finally, the head generates the model's predictions based on the combined features. The 2SYOLO model processes images by first passing them through multiple convolutional layers in the backbone. The extracted features from backbone are then forward to the neck, which enhances multi-scale feature representation. Decision making for classification, bounding block regression and score prediction is done by head.

C. 2D Face Descriptor with Deep Learning

Convolutional neural networks (CNNs) were developed for more than two decades beginning with the introduction of LeNet architecture [11]–[13]. CNNs are a special type of multi-layer neural networks in which convolution operations are used to extract distinctive features from input images. Therefore, raw images are not fed directly from one layer to the next, but are transformed by convolution operations before being forward. Extensive empirical evidence shows that deeper CNNs architectures produced greater performance [11], [12]. CNNs are composed of multiple layers including feature extractors (e.g., convolution operation), dimensionality reduction (e.g., max pooling) and classification layers (e.g., softmax layer) [7]. The cropped and aligned face images resulting from the previous section are color images with the size of $W \times H$, where $W = 130$ and $H = 200$ in our implementation. Therefore, input image is represented as a tensor $I \in \mathbb{Z}^{W \times H \times C}$ and is fed to the CNNs with shape $W \times H \times C$, where W and H are width and height of an image and C is the number of channels, i.e. $C = 3$ for RGB images. Normally, the outputs of CNNs are produced by softmax layer. One can define CNNs as a K class probability distribution where the output is maximized at index k^{th} if and only if k^{th} class is identical to the identified object. Let us define by v_j the target of CNNs classification where j is the index of class j^{th}. By minimizing [11]–[13]

$$\left\|\sigma(\mathbf{y})_j - v_j\right\|; \forall j \in K, \tag{1}$$

for K class problem, the CNNs is converged. $\sigma(\mathbf{y})_j$ is a softmax probability and can be described by [12]

$$\sigma(\mathbf{y})_j = \frac{e^{y_j}}{\sum_{k=1}^{K} e^{y_k}}; j = 1, \ldots, K. \tag{2}$$

y_j is a fully connected layer where all neurons from the previous layer are connected with weights to the next layer, hence the term fully connected (FC). Let us define by $FC(x)$, the output of the fully connected layer with [11]

$$FC(x) = a(Wx + b) \in \mathbb{R}^m. \tag{3}$$

where a is an activation function. The layer $FC(x)$ is fully connected with n inputs from the previous neuron layer and m output dimensions, computed using a trainable weight matrix $W_{m \times n} \in \mathbb{R}^2$ and bias vector $b \in \mathbb{R}^m$. Typically, the activation function a used in CNNs is a nonlinear rectified linear unit (ReLU) which is defined by [13]

$$a(u) = \max(0, u), u \in \mathbb{R}, \tag{4}$$

It is easy to prove that $a'(u) = 1$ for $u > 0$ and that $a'(u) = 0$ for $u < 0$. ReLU has an advantage over traditionally sigmoid function in that it promotes faster convergence. This is primarily because the sigmoid function may lead to the vanishing gradient problem when the input u is far from zero [12], i.e., $u \to \infty$ or $u \to -\infty$. The pooling layer is used to reduce the dimensionality of feature maps which are generated from conv and ReLU layers. In this paper, max pooling is used for selecting maximum value within each spatial neighbor with kernel size of 2 and the stride of 2. This leads to downsampling the feature maps, which helps CNNs facilitates faster convergence in training phases.

In the face authentication system, we need to create a face template and store it in ID card or in database. The face template (face descriptor) will be used in face matching process where the matching may be 1-to-1 or 1-to-many. However, the output of CNNs produced by softmax layer is not appropriate for this case and can not be used as a face template. We proposed here to

use the output from several FCs layer as a 2D face descriptor (2DFD). Small feature representation can be used for faster classification with the risk of weak identity. We need templates for face authentication where the templates of the same class must be similar while those of different individuals remain distinct. Based on our experiments, more and more FCs layers can be used to construct a 2D face descriptor (2D-FD) and helps us to improve the overall accuracy of face authentication system. The 2DFD serves as a face template for face matching (face authentication in our case). Our proposed architecture is shown in Fig. 3.

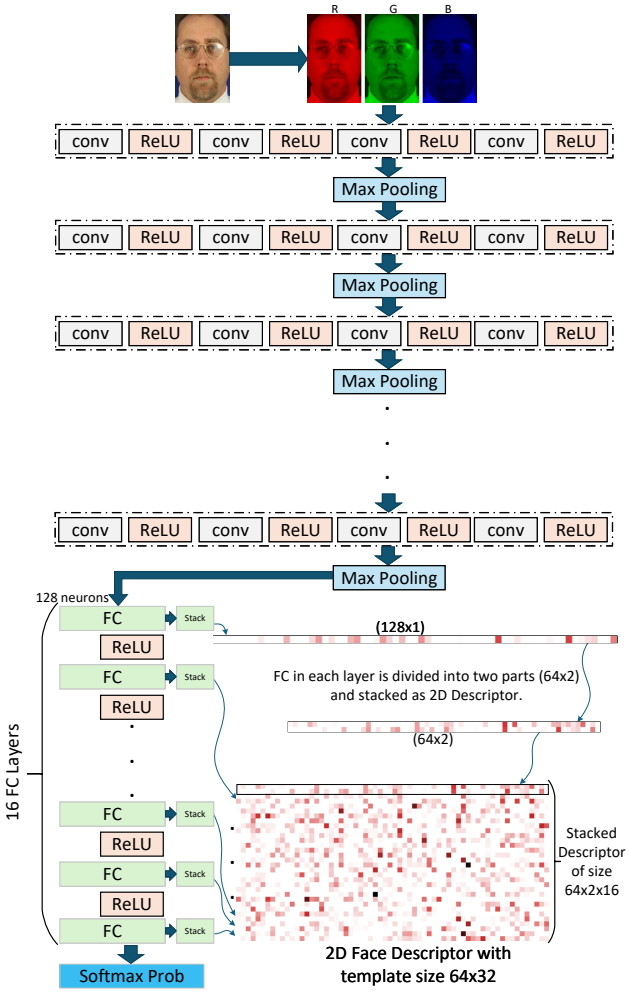

Fig. 3: The architecture of our 2DFD is based on fully connected layers from CNNs. The 2DFD is formed by stacking multiple FC layers, enabling the extraction of more distinctive features. It should be noted that the color in 2DFD is artificially created for illustrative purposes only.

The details of our CNNs architecture in Fig. 3 can be described as follows. First, the input color image is separated into three channels Red, Green and Blue (R-G-B images). The R-G-B image channels are then fed into a series of multiple convolution layers (conv) and followed by ReLU to extract features. In order to preserve salient features while reducing redundancy, those features from conv and ReLU are then processed by the max pooling layers (Max Pooling). We repeat these conv, ReLU and Max Pooling for several layers to produce more and more salient features before passing them to the next layer. We then construct a sequence of the fully connected layers (FC) each with 128 neuron nodes followed by ReLU. The result of each ReLU is then fed to the next FC and ReLU for 16 layers. The output of each FC with 128 neuron nodes is split equally into two halves of size 64×2 which are stacked to form a 2D descriptor. As a result, the final 2D descriptor has the size of 64×32, i.e. $64 \times 2 \times 16 = 2048$, equivalent to 2KB. This compact size of 2DFD is good for efficient storage and fast processing, while still strong enough to serve as discriminative face descriptor. Let us define by $\lambda_i^k(r,c) \in \mathbb{R}$ the 2D face template for class k^{th} of image i^{th} where $r \in R$ and $c \in C$, i.e. $R = 32$ and $C = 64$, respectively. In order to standardize the face template, we normalize it by $\hat{\lambda}_i^k(r,c) = \frac{\lambda_i^k(r,c) - min}{max - min}$ so that $\hat{\lambda}_i^k(r,c) \in [0,1]$. For illustration proposes, $\hat{\lambda}_i^k(r,c)$ should be scaled to gray level range by subtracting and multiplying it by 255, i.e., $\hat{\lambda}_i^k(r,c) := 255 - (\hat{\lambda}_i^k(r,c) * 255)$. Additionally, the color of the 2DFD shown in Fig. 3 is artificially filled to enhance visual understanding. One can observe the difference between 1D (the 128×1 1D descriptor) and 2D (the 64×2 and 64×32) descriptors that the 2D descriptor exhibit more structured patterns compared to the 1D vector. Hence, the between-class variances can be maximized while maintaining the with-in class variances as low as possible. This face template $\hat{\lambda}_i^k$ will be used to construct the volume deep face in the next section.

D. Volume Deep Face

In this section, we propose the volume deep face (VDF) which is constructed from 2DFD $\hat{\lambda}_i^k(r,c)$. Let us suppose that we generate multiple 2DFDs $\hat{\lambda}_i^k(r,c), i = 1, ..., N$ where N is the number of training images for class k^{th}. The VDF is shown in Fig. 4. This VDF is a 3D representation inheriting from multiple 2DFDs. Therefore, more distinctive features can be achieved. The VDF can be regarded as a tensor-valued function and formulated as

$$\Lambda = \hat{\lambda}_1^k \oplus \hat{\lambda}_2^k \oplus \cdots \hat{\lambda}_i^k \cdots \oplus \hat{\lambda}_n^k, 1 \leq i \leq n \qquad (5)$$

where $\oplus$ is a concatenate operation. Therefore, each $\hat{\lambda}_i^k$ is concatenated in the direction of index i. n is the number of feature maps generating from FC layer of CNNs. In summary, given input images from class k^{th}, create multiple 2DFDs $\hat{\lambda}_i^k$ from FC layer, then concatenating all features to form the VDF Λ. It should be noted that the VDF is generated from CNNs where the learnable weight W has been trained with error minimization. The details of generating VDF is summarized in algorithm 1.

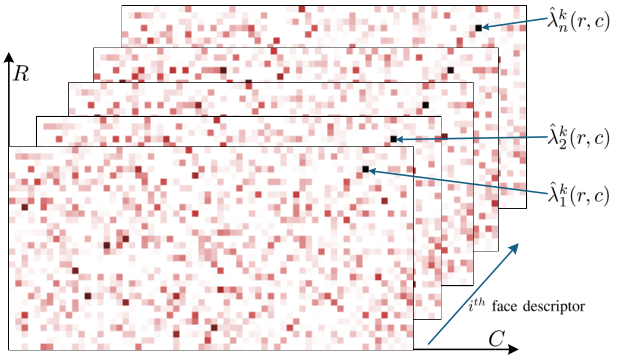

Fig. 4: The volume deep face which is constructed from multiple 2DFDs $\hat{\lambda}_i^k(r,c)$ where i is an index of i^{th} face descriptor generating for class k^{th} with the size $R \times C$.

Algorithm 1 Construction of volume deep face (VDF).

Let $I_i^k \in \mathbb{Z}^{W \times H \times C}$ be the training image i^{th} of class k^{th}.
repeat
  for each $k \in K$ **do**,
    for each $i, 1 \leq i \leq n$ **do**,
      compute conv and ReLU from I_i^k
      apply max pooling to ReLU
      $FC(x)^j \leftarrow a(Wx+b) \in \mathbb{R}^m, 1 \leq j \leq 16$
      stack $FC(x)^j \ \forall j$ to generate $\hat{\lambda}_i^k$
    end for
    construct Λ^k for each class k^{th}
  end for
until all classes K have been generated

E. Volume Matching

Let Λ_A be the VDF of images generating by algorithm 1 and let $\hat{\lambda}^t(r,c)$ denote the 2DFD of a test image, as described in the previous section. As our VDF can be generated from multiple images in the training set, we assume that during testing phase, only a single image may be used per test. In such cases, images in the training set can be used to constructed the VDFs, while in the test phase only 2DFD will be generated. We measure the similarity between the VDF and 2DFD by

$$\gamma = \frac{1}{L}\sum_{i=1}^{n}\sum_{r=1}^{R}\sum_{c=1}^{C}\left\|\hat{\lambda}_i^A(r,c) - \hat{\lambda}^t(r,c)\right\|_2, \ \forall \hat{\lambda}_i^A \in \Lambda_A \tag{6}$$

where $L = n \times R \times C$ is the total number of elements in the VDF. If $\hat{\lambda}^t(r,c)$ is identical to the $\hat{\lambda}_i^A(r,c)$, then $\gamma \to 0$ otherwise $\gamma \to \infty$

III. EXPERIMENTAL RESULTS

The images used in this paper were collected from the standard XM2VTS face database [6] as shown in Fig. 5. The database has 295 subjects each of which captured by 8 shots with 4 distinct sessions during 4 months resulting in a total of 2,360 images. The database was divided into three sets: training set, evaluation set and test set. [6].

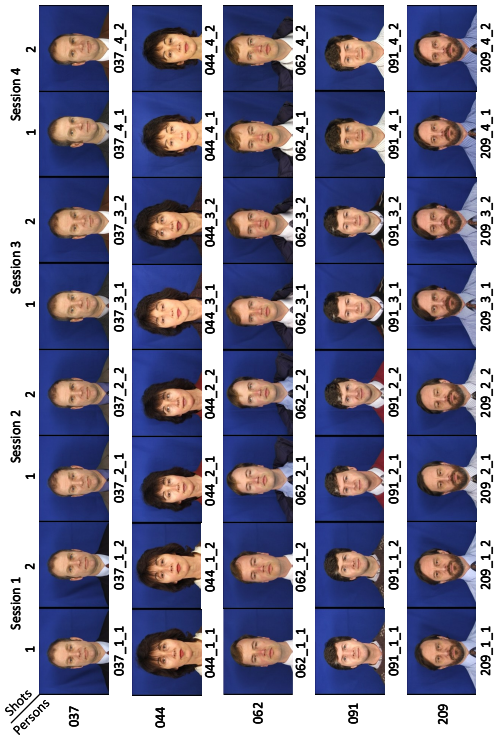

Fig. 5: Examples of images in XM2VTS face database used for face authentication system.

The corresponding 2DFDs for each cropped and aligned face image were generated and depicted in Fig. 6. Columns (a) and (c) are examples of aligned and cropped face images, while columns (b) and (d) represent the corresponding 2DFDs of (a) and c), respectively. It was cleared that the 2DFDs of the same individual are highly similar in which the within-class variance is minimized. Moreover, the 2DFDs of different classes show significant differences, which helps maximize the between class variance. We can conclude that the micro pattern on each element (r,c) of 2DFD $\hat{\lambda}$ represents the 2D structure generated from the face image exhibiting more discriminative features for classification.

A. Face Verification

The performance of our proposed method was evaluated based on the protocol described in [6]. False acceptance (FA) occurs when an imposter (false identity) is incorrectly accepted as a client (true identify). In contrast, false rejection (FR) occurs when the true identity has been rejected. The FA and FR were measured by $FA = \frac{EI}{I} \times 100$ and $FR = \frac{EC}{C} \times 100$, where EI and EC represent the number of imposter acceptances and client rejections, respectively. I and C denote the total number of imposter and clients claims in the system. In

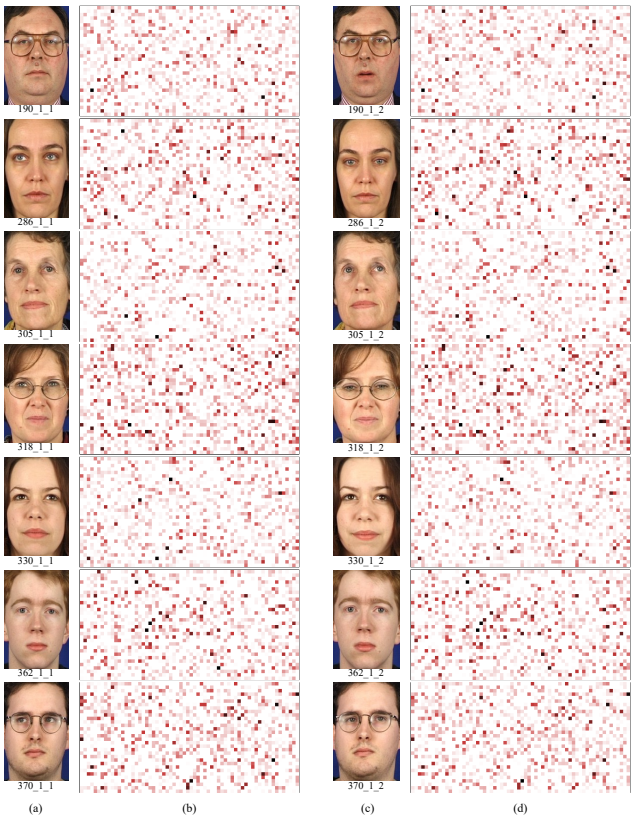

Fig. 6: The 2DFDs generated from images in XM2VTS database (a) face images from session 1 shot 1, (b) the corresponding 2DFDs of (a), (c) face images from session 1 shot 2, (d) the corresponding 2DFDs of (c).

this paper, $I = 112000$ and $C = 400$. Table I shows the error rates of our proposed method in comparison to the other approaches. For fair comparison, all methods in Table I were evaluated on identical XM2VTS inputs under the same protocol [6]. We obtain the total error rate (TER) with 0.01161% which is the lowest error rate, where $TER = FA + FR$. It is worth noting that, with a very low error rate of our proposed method, implementing a face authentication system for airport check-in process is feasible.

TABLE I: Error Rates on XM2VTS Database. (The bold values indicate the best performance.)

Methods [6]	**Test Set**		
	FA (%)	**FR** (%)	**TER** (%)
UniS-ICPR2000	2.30	2.50	4.80
IDIAP-Marcel	1.748	2.0	3.75
IDIAP-Cardinaux	1.84	1.50	3.34
MUT-UniS-STT	0.97	0.50	1.47
UCL	1.71	1.50	3.21
TB	5.61	5.75	11.36
UniS-ECOC	0.86	0.75	1.61
UniS-NC	0.48	1.00	1.48
weighted DeepFace [8]	0.01429	0.0	0.01429
Our Proposed Method	**0.01161**	0.0	**0.01161**

IV. CONCLUSIONS

We have proposed a novel face representation, the Volume Deep Face (VDF), for a face authentication system. The VDF is composed of multiple concatenated 2D face descriptors in the tensor form allowing us to generate more discriminative features of face representations. The 2DFDs were derived from multiple fully connected layers of CNNs that has been pre-trained prior to the VDF construction process. Additionally, input of face images were aligned and cropped with our proposed two-stage YOLO based face and eye detection which helps us increasing the performance of the overall system. We obtained the total error rate with 0.01161% which is the lowest error reported to date on the XM2VTS database. Our proposed method can be applied to face verification in the airport check-in process.

REFERENCES

[1] D. Ying., "Exploring PCA-based feature representations of image pixels via CNN to enhance food image segmentation," 10.48550/arXiv.2411.01469., 2024.
[2] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720, July 1997.
[3] M. Ghorbani, A. T. Targhi and M. M. Dehshibi, "HOG and LBP: Towards a robust face recognition system," 2015 Tenth International Conference on Digital Information Management (ICDIM), Jeju, Korea (South), 2015, pp. 138-141.
[4] T. Ahonen, A. Hadid and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037-2041, Dec. 2006.
[5] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701-1708, 2014.
[6] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sandersonan, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi and N. Mavity, "Face Verification Competition on the XM2VTS database," in Proc. of 4th Int. Conf. Audio and Video Based Biometric Person Authentication, Lecture Notes in Computer Science, Vol. 2688, Springer-Verlag, pp. 964-974, 2003.
[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," arXiv preprint arXiv:1408.5093, 2014.
[8] –, "Robust Face Recognition based on weighted DeepFace," 2017 International Electrical Engineering Congress (iEECON), Pattaya, Thailand, 2017, pp. 1-4, doi: 10.1109/IEECON.2017.8075885.
[9] C. Li, et. al., "YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications," Technical Reports, 2022, https://arxiv.org/abs/2209.02976.
[10] J. Yangqing, et. al, "Caffe: Convolutional Architecture for Fast Feature Embedding," arXiv preprint arXiv:1408.5093.
[11] Y. LeCun, et. al, "Handwritten Digit Recognition with a Back-Propagation Network," Advances in Neural Information Processing Systems, vol. 2, 1989.
[12] I. Goodfellow, Y. Bengio and A. Courville, "Deep Learning," MIT Press, 2016.
[13] N. Buduma, N. Buduma and J. Papa, "Fundamentals of Deep Learning, "O'Reilly Media, Inc., 2022.