ІНФОКОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА ЕЛЕКТРОННА ІНЖЕНЕРІЯ
INFORMATION AND COMMUNICATION TECHNOLOGIES, ELECTRONIC ENGINEERING

№ 4 (2), 2024

RESEARCH ON THE STATE-OF-THE-ART DEEP LEARNING BASED MODELS FOR FACE DETECTION AND RECOGNITION

A. Sydor[1][ORCID: 0009-0000-7838-8189], D. Balazh, Yu. Vitrovyi[2], O. Kapshii, O. Karpin[3], T. Maksymyuk [2] [ORCID: 0000-0002-2739-9862]

[1]LLC «EUROSOFTWARE-UA», 79008, Ukraine, Lviv region, Lviv, 33a Lychakivska str.
[2]Lviv Polytechnic National University, 12 S. Bandery str., 79013, Lviv, Ukraine
[3]Infineon Technologies, 20 Luhanska str., 79034, Lviv, Ukraine

Corresponding author: A. Sydor (e-mail: artur.sydor.mitpa.2022@lpnu.ua).

(Submitted on 1 June 2024)

The problem of building a face recognition pipeline faces numerous challenges such as changes in lighting, pose, and facial expressions. The main stages of the pipeline include detection, alignment, feature extraction, and face representation. Each of these stages is critically important for achieving accurate recognition. The article analyzes and compares modern algorithms and models for face detection and recognition in terms of their ability to correctly identify true positives (TP) and true negatives (TN) while minimizing false negatives (FN) and false positives (FP) in facial recognition. Classical algorithms and lightweight models, such as MediaPipe, offer the highest speeds but sacrifice some accuracy. Conversely, heavier models like RetinaFace deliver greater accuracy at the expense of speed. For systems prioritizing maximum detection accuracy and minimizing missed faces, models like DSFD or RetinaFace-Resnet50 are recommended, despite their slow performance and unsuitability for real-time detection. If the primary goal is maximum detection speed and occasional missed faces in uncontrolled conditions are acceptable, an SSD face recognition solution is preferable. For applications requiring a balanced approach to speed and accuracy, the RetinaFace-MobilenetV1 model is optimal in terms of real-time detection speed and satisfactory accuracy. The ArcFace model demonstrates superior performance with a TP rate of 0.92 and a TN rate of 0.91, indicating a high accuracy in both identifying the correct person and rejecting mismatched images. ArcFace also maintains a low FP rate of 0.09. FaceNet follows with a TP rate of 0.89 and an impressive TN rate of 0.94, showcasing its proficiency in avoiding incorrect matches. In contrast, VGGFace, DeepFace, and OpenFace show moderate TP rates between 0.61 and 0.78, coupled with higher FN and FP rates. The DeepID model exhibits the lowest performance, with a TP rate of 0.47 and a TN rate of 0.60, reflecting substantial difficulties in accurate identification. The conclusions emphasize the importance of selecting models based on accuracy, speed, and resource requirements, suggesting RetinaFace and ArcFace/FaceNet as good trade-off options.

Keywords: *face detection, face recognition, convolutional neural networks, feature extraction*.
UDC: 621.3

1. Introduction

Building a face recognition pipeline faces numerous challenges that must be addressed for successful implementation. Achieving high accuracy and practicality in face recognition is critically important. Issues such as changes in lighting, pose, and facial expressions can affect the system's ability to correctly identify individuals. The quality of training data significantly impacts the performance of face recognition models.

Biases in data, such as insufficient representation of certain demographic groups, can lead to skewed and less reliable results, impacting fairness. The face recognition pipeline consists of several key stages, each playing a crucial role in the outcome. Face detection identifies and determines the location of faces in an image or video frame, which is fundamental [1]. Errors at this stage can result in missed faces or false positives, affecting subsequent steps. Face alignment normalizes the orientation and position of the face for consistent feature extraction. Proper alignment enhances the model's ability to extract meaningful features, improving recognition accuracy [2]. Feature extraction involves isolating relevant facial features that distinguish one face from another. The effectiveness of feature extraction directly influences the pipeline's performance. Extracting distinctive features improves recognition accuracy. Face representation (embedding) transforms facial features into a compact numerical representation (embedding) suitable for comparison [3]. Well-crafted embeddings allow for precise differentiation of faces during verification or recognition. The aim of this article is to identify the methods and models necessary for constructing an efficient face recognition pipeline that achieves high accuracy, flexibility, and seamless integration with various external systems.

2. Related work

2.1 Advances in face detection technologies

In recent years, face recognition systems have gained immense popularity due to their integration across various sectors. With the advancement of machine learning and deep learning methods, face detection and recognition models are increasingly used to achieve more accurate results.

The article [4] outlines the fundamental concepts of convolutional neural networks (CNNs), detailing the necessary layers for their construction and the optimal way to structure the network for most image classification tasks. CNNs outperform other deep learning methods for computer vision systems, providing superior performance in image recognition and even surpassing human accuracy in some cases.

The study [5] examines the Viola-Jones algorithm, the first real-time face detection system. It achieves fast and accurate detection through three components: an integral image for feature calculation, the Adaboost method for feature selection, and a cascade of attention for efficient computational resource allocation. Since the Viola-Jones algorithm typically provides multiple detections, a post-processing step is proposed to reduce detection redundancy using a robust argument.

In the work [6], the complex task of simultaneous localization and alignment of faces of various scales in images is explored. The first one-step solution called RetinaFace is proposed. This method outperforms existing state-of-the-art methods in the most challenging face detection tests to date. Moreover, combining RetinaFace with advanced face recognition practices immediately improves accuracy.

The study [7] discusses a method for feature extraction from face images based on a deep learning algorithm. The article addresses both local and global feature extraction based on deep learning and proposes a classification training method based on a deep model that combines the local pattern and the GLQP feature extraction algorithm. The results demonstrate that deep learning has a greater automated feature extraction capability than shallow learning. The development of a rich environment function ensures a sufficient number of samples for training the deep model, fostering the development of more unique features.

Given the current diversity of models with varying complexity for face detection and recognition, further research is necessary to build a face recognition pipeline with high speed and accuracy.

This section provides a comprehensive analysis of face detection methods, examining both classical techniques and the limitations inherent to them. The evolutionary trajectory of face detection is meticulously traced, covering traditional methods and transitioning to the modern landscape dominated by deep learning algorithms. The section also describes the current state of face recognition tasks, emphasizing the crucial role of effective representation through embeddings as the primary means of addressing challenges in face recognition. Various models for creating such embeddings are considered, and their effectiveness is compared.

2.2. Classic Face Detection Algorithms

Haar Cascades. One of the earliest face detection algorithms is the Viola-Jones algorithm. Working with grayscale images, the Viola-Jones algorithm interprets images as a set of Haar features represented by light and dark rectangles. There are many variations of Haar features with different arrangements of light and dark areas within a rectangle. Each feature is evaluated by the first classifier. If it rejects the feature (indicating an area that does not contain a face), it is immediately discarded. If the feature is accepted (indicating the presence of a face), it moves to the next classifier. Essentially, the Viola-Jones detector transforms the task from direct face detection to systematic exclusion of non-face areas. This cascading approach proves highly effective for fast face detection. While it significantly improved detection speed and accuracy, it had limitations and struggled with detecting faces in noisy images. Over the years, many improvements have been made. The Haar Cascade algorithm has been used not only for face detection but also for detecting eyes, license plates, and more [5].

DLib-HOG. The widely used face detector DLib-HOG employs the classic Histogram of Oriented Gradients (HoG) feature in combination with a linear classifier, an image pyramid, and a sliding window detection scheme. This method is known for its effectiveness in detecting faces by analyzing the gradients of image intensities, which helps in identifying the edges and contours of faces. DLib-HOG uses five specific HoG filters to improve detection accuracy: one for frontal faces, one for faces turned to the left, one for faces turned to the right, one for frontal faces rotated to the left, and one for frontal faces rotated to the right. This multi-filter approach allows DLib-HOG to account for various orientations of the face, making it versatile in different scenarios. The image pyramid technique scales the image to multiple resolutions, ensuring that faces of different sizes can be detected effectively. The sliding window detection scheme systematically scans the image, applying the HoG filters at each step to locate faces. Although DLib-HOG was revolutionary at its inception and remains widely used due to its simplicity and relatively low computational cost, it has largely been superseded by more advanced deep learning methods. These modern approaches offer greater accuracy and robustness, particularly in diverse and challenging conditions where traditional methods may struggle. The advanced deep learning models, such as those employing convolutional neural networks (CNNs), can learn more complex features and patterns in the data, enabling them to handle variations in lighting, pose, and occlusion more effectively. In the subsequent sections, we will delve into these modern approaches, exploring their architectures, training techniques, and the specific capabilities that make them superior to classical algorithms like DLib-HOG [1].

2.3. Deep Learning-Based Face Detection Algorithms

Classical face detection methods may struggle to detect faces in multiple frames, which can lead to the program not functioning as desired or causing complications in the system. Even if faces are detected in every frame, the process can be time-consuming. This slows down the application and sometimes compromises its integrity. Modern face detectors provide high accuracy (ensuring no face goes undetected) at very high speeds and can also be used in microprocessors with low computational power.

SSD (Single Shot MultiBox Detector). The SSD model detects objects in a single pass through the input image, unlike other models that pass through the image multiple times to generate detection outputs. The SSD model consists of two parts. First is the base model, which is a typical pre-trained image classification network that works as a feature extractor, without last dense layer. The second is the SSD head, which is composed of several convolutional layers stacked on top of the base model. This provides the output in the form of bounding boxes around objects. These convolutional layers detect various objects in the image [8].

MTCNN (MultiTask Cascaded Convolutional Neural Network). MTCNN is a state-of-the-art tool for face detection that uses a three-stage neural network detector. The first step involves scaling the image to different sizes to create an image pyramid. This image pyramid serves as the input for the next stage. The first stage includes a fully convolutional network (FCN) called the Proposal Network (P-Net), designed to obtain candidate windows and regression vectors for the bounding boxes. After obtaining the regression vectors, a refinement process is applied to merge overlapping areas. The final result of this stage is all candidate windows after refinement, reducing their number. The Refinement Network (R-Net) receives all candidates created by the Proposal Network (P-Net). This network is also a convolutional neural network (CNN). R-Net further reduces the number of candidates, refines the positions of the bounding boxes through regression, and uses non-maximum suppression (NMS) to merge overlapping candidates. The R-Net output consists of three components: binary classification indicating whether a face is present in the input image, a 4-element vector for the bounding box around the face, and a 10-element vector for facial landmark localization. This stage is similar to R-Net, but the Output Network (O-Net) provides a more detailed description of the face and determines the positions of five facial landmarks, including the eyes, nose, and mouth. MTCNN is one of the most accurate face detection algorithms, especially in challenging conditions like poor lighting and partial occlusion of the face. It is highly robust to changes in facial appearance, such as variations in pose, expression, and makeup. Additionally, MTCNN offers relatively high operational speed (with GPU), making it a good choice for real-time video processing where speed is crucial. Since convolutional neural networks (CNNs) process images in RGB format, MTCNN utilizes color information effectively. However, MTCNN has a relatively complex structure, which can complicate its implementation and optimization. It also requires significant computational resources, which can limit its use on mobile devices and other platforms with limited computational capabilities [9].

Dual Shot Face Detector (DSFD) is an innovative approach to face detection that addresses three main aspects of face recognition. Firstly, it includes a Feature Enhancement Module (FEM), which enhances the initial feature maps, extending the single-shot detector to a dual-shot detector. This module helps incorporate information from the current layer along with feature maps from previous layers, maintaining the contextual relationship between anchors. This results in more distinct and robust features. Secondly, DSFD includes Progressive Anchor Loss (PAL), which is calculated with two sets of anchors. Smaller anchors are used in the first shot and larger ones in the second shot, effectively ensuring feature development. Thirdly, DSFD employs Improved Anchor Matching (IAM), which includes anchor-based data augmentation. This ensures better matching between anchors and actual information, leading to better initialization of the face bounding box regressor. All these aspects are interchangeable and can work simultaneously to improve performance. As a result, DSFD remains robust even under changes in lighting, pose, scale, and occlusion [10].

RetinaFace is a cutting-edge deep learning model for face detection, performing three different tasks related to face localization: face detection, 2D face alignment, and 3D face reconstruction, all based on a one-stage framework. It utilizes three main elements: the feature pyramid, the one-stage method, and context modeling. The feature pyramid takes a single image as input and generates feature maps at different scales, which has proven to be a key tool for improving object detection in recent years. The one-stage method simplifies the process by requiring only one pass through the network to generate object bounding boxes, increasing efficiency. Context modeling involves learning contextual information from images using deformable convolutional networks (DCNs), which allows for more adaptive operations at different scales of object features. RetinaFace's feature pyramid network (FPN) is responsible for extracting features from five different levels of the 2D image. The first four feature maps are computed using a pre-trained ResNet model, while the smallest feature map at the top is obtained by 3x3 convolution with a stride of 2. The context module head with five different filters extracts additional contextual information from these features, which is then directed to the multi-level loss function. The loss function used by RetinaFace includes several components. Face Classification Loss is a softmax loss for binary classes (face/no face). Face Bounding Box Regression Loss normalizes and represents the target bounding boxes in the format [(x_center, y_center, width, height)]. Facial Landmark Regression Loss also normalizes the target values. Dense Regression Loss uses supervised learning signals to enhance the importance of accurate face and landmark localization. These loss components collectively improve the accuracy of face and landmark detection, making RetinaFace an effective method for solving face detection and recognition tasks. RetinaFace is known for its exceptional accuracy in face detection, achieving 91.4% average precision (AP) on the WIDER FACE dataset, even in challenging conditions such as low lighting and occlusion. It is robust to changes in facial appearance, including different poses, facial expressions, and makeup. RetinaFace performs three different face detection tasks simultaneously: face detection, 2D facial landmark localization, and 3D face reconstruction. However, RetinaFace operates slower than some other face detection algorithms and requires significant computational resources, which can limit its use on mobile devices and other platforms with limited resources [7].

2.4. Main Problems and Limitations of Face Detection Algorithms

The process of face detection is complicated by several factors. One significant challenge is glare, which can severely limit the capability of any face detection system. When only a portion of the face is visible due to glare, it becomes difficult to accurately determine the presence of a face in the frame. Lighting conditions also pose a major problem for face detection. Changes in lighting can obscure facial features, leading to inaccurate detections, especially if the system is not designed or trained to handle lighting variations. Skin color presents another challenge. Some face detectors show bias toward certain skin colors, and different skin tones can behave differently under varying lighting conditions. This variability adds complexity to the detection process. The pose and orientation of the face are crucial factors. Many detection methods are optimized for frontal faces and struggle with faces that are turned sideways or slightly angled. This limitation affects the detector's performance in real-world scenarios where face orientation can vary greatly. Facial expressions must also be considered. In the real world, faces rarely remain neutral. Changes in facial expressions alter facial features, which can confuse the detection system if it is not trained to recognize these variations. Accessories and facial changes, such as sunglasses, face masks, beards, tattoos, and heavy makeup, can obscure facial features and impede accurate detection. These elements need to be accounted for in the design and training of the face detection system. Lastly, the scale of the face relative to the image or video frame can vary. Faces may be too small to detect if the system is not capable of identifying faces at different scales. Effective detectors must handle various face sizes to be useful in diverse applications. Addressing these challenges is essential for developing reliable face detection systems that perform accurately under a wide range of conditions and scenarios [11].

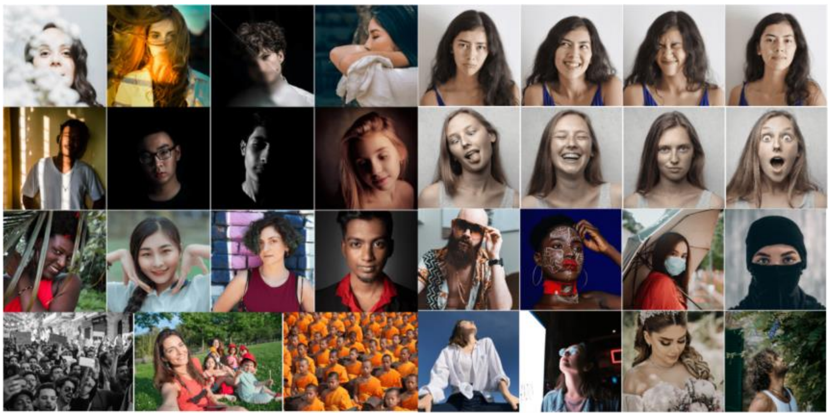

Fig. 1 Examples of faces with difficult detection conditions

3. Experimental Evaluation of the Face Detection and Recognition Models

3.1. Numerical Evaluation of Face Detection Models

Understanding the metrics used in face detection requires first understanding how they differ from metrics used in other types of machine learning tasks, such as classification and regression. Object detection, which includes face detection, is more complex because it involves both identifying the presence of objects and precisely locating them within an image. In face recognition tasks, metrics like accuracy, precision, recall, and the F1 score evaluate the model's ability to correctly classify instances into predefined categories.

As seen in Figure 2, the Haar Cascades algorithm effectively detects faces of various sizes but struggles with faces with accessories or makeup. It also has good performance speed (0.818 seconds). The DLib-HOG algorithm performs well with frontal faces but has difficulties detecting small faces. However, its advantage is high-speed performance (0.155 seconds). The SSD algorithm detects faces in the foreground effectively but has trouble with detecting smaller, distant faces. Its significant advantage is the highest speed (0.055 seconds) among the detectors studied in this work. The MTCNN algorithm detects faces of various sizes well and does not struggle with faces with accessories or makeup, but its speed (3.895 seconds) is not high. The DSFD algorithm detects faces of various sizes well and does not have issues with detecting very small, distant faces, but its disadvantage is the lowest speed (47.74 seconds) among the detectors studied. The RetinaFace algorithm detects faces of various sizes well and does not struggle with faces under glare. It also has high-speed performance (0.693 seconds) (Table 1).

Table 1

Comparison of Detectors by Face Detection Time

Model	Time, s
Haar Cascades Хаара	0.818
DLib-HOG	0.155
SSD	0.055
MTCNN	3.895
RetinaFace MobilenetV1	0.693
Dual Shot Face Detector	47.74

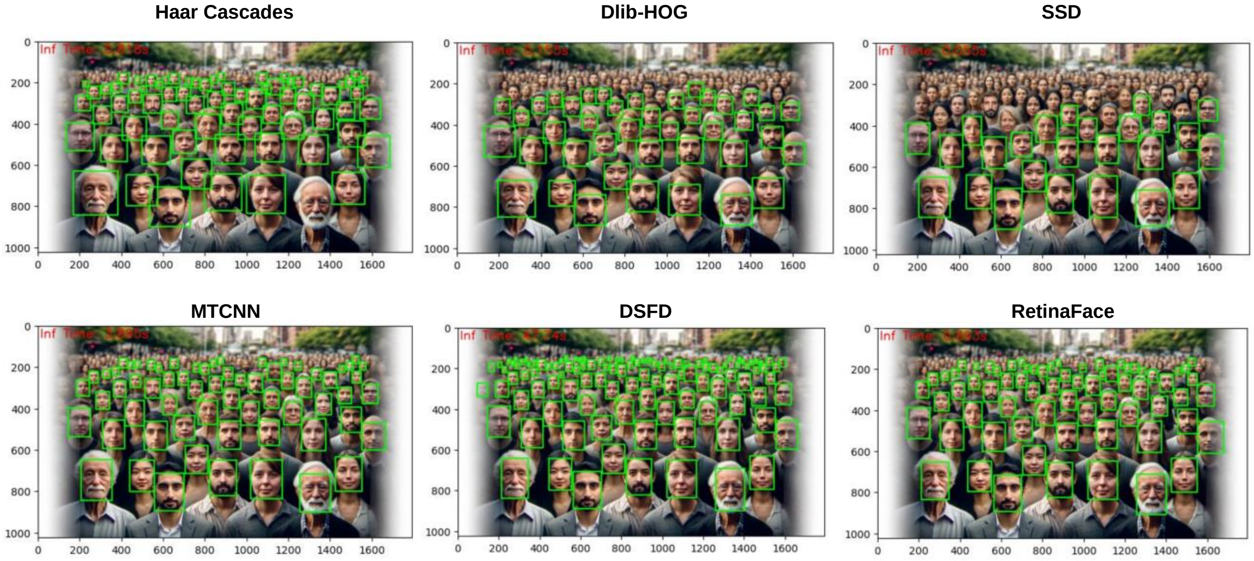

Fig. 2. Experimental results of the various face detection models on the test image

From Table 1, it is evident that the SSD model is the fastest for face detection, with an execution time of only 0.055 seconds. DLib-HOG is also very fast with a time of 0.155 seconds. Classical algorithms like Haar Cascades (0.818 seconds) and lightweight models like RetinaFace MobilenetV1 (0.693 seconds) demonstrate acceptable speeds. However, more advanced models, such as the Dual Shot Face Detector (47.74 seconds), are significantly slower. In summary, there is a trade-off between accuracy and speed. Classical algorithms and lightweight models like MediaPipe achieve the highest speeds but with lower accuracy, while heavier models like RetinaFace operate slower but more accurately. If the system needs to achieve the highest detection accuracy and there is a requirement not to miss any faces, models like DSFD or RetinaFace-Resnet50 should be chosen. However, such systems will be very slow and not suitable for real-time detection. If the system needs to achieve the maximum detection speed and missing faces in uncontrolled conditions is not a problem, then an SSD face recognition solution should be selected. If the face detection system requires a good balance of speed and performance, the RetinaFace-MobilenetV1 model should be considered. This model is very fast with real-time detection speeds while still providing acceptable accuracy.

3.2. Numerical Evaluation of Face Recognition Models

A comparison was conducted to evaluate the efficiency of several popular face recognition models, including VGG-Face, FaceNet, OpenFace, DeepFace, DeepID, and ArcFace.

VGG-Face is one of the most widely used image recognition models based on deep convolutional neural networks. The VGG architecture became well-known for its high results in the ImageNet challenge. Developed by researchers at the University of Oxford, VGG-Face has a structure similar to the standard VGG model but is fine-tuned specifically for facial recognition. This tuning allows it to handle various facial features and expressions effectively, making it highly reliable for diverse facial recognition tasks [12].

Google FaceNet, developed by Google researchers, is considered a state-of-the-art model for face detection and recognition using deep learning. FaceNet stands out because of its ability to perform not just face recognition but also verification and clustering. The model can group photos of people with the same identity, which is particularly useful in managing large photo libraries. Its architecture efficiently encodes facial features into a compact representation, facilitating high performance in both accuracy and speed [13].

Developed by researchers at Carnegie Mellon University, OpenFace is heavily inspired by the FaceNet project. However, OpenFace is designed to be lighter and comes with a more flexible licensing arrangement, making it accessible for various applications. Despite its lighter framework, OpenFace maintains robust facial recognition capabilities, leveraging deep learning to process and identify facial features efficiently [14].

Facebook DeepFace, developed by Facebook researchers, uses a deep neural network with nine layers. This model was trained on a massive dataset of four million faces belonging to over 4,000 individuals, making it one of the largest face datasets at the time of its release. The extensive training allows DeepFace to recognize faces with a high degree of accuracy, approaching human-level performance. The model's robustness comes from its ability to handle a wide range of facial variations and conditions, such as lighting and angles [15].

The DeepID face verification algorithm, developed by researchers at the Chinese University of Hong Kong, was one of the pioneering models to use convolutional neural networks for facial recognition. DeepID-based systems were among the first to surpass human performance in this task. The model focuses on extracting high-dimensional facial features, which helps in distinguishing between similar faces. Its development marked a significant advancement in the field, demonstrating the potential of deep learning for complex recognition tasks [16].

ArcFace, the latest model jointly developed by researchers from Imperial College London and InsightFace, introduces innovative techniques in facial recognition. ArcFace employs an additive angular margin loss to enhance the discriminative power of the face recognition system. This method improves the model's ability to differentiate between faces with subtle differences, making it highly effective for applications requiring precise facial identification. The collaboration between academic and industry experts has resulted in a model that pushes the boundaries of what facial recognition technology can achieve [17].

The results of the model comparisons are presented in Tables 2 and 3, and performance graphs for each model are shown in Figure 3.

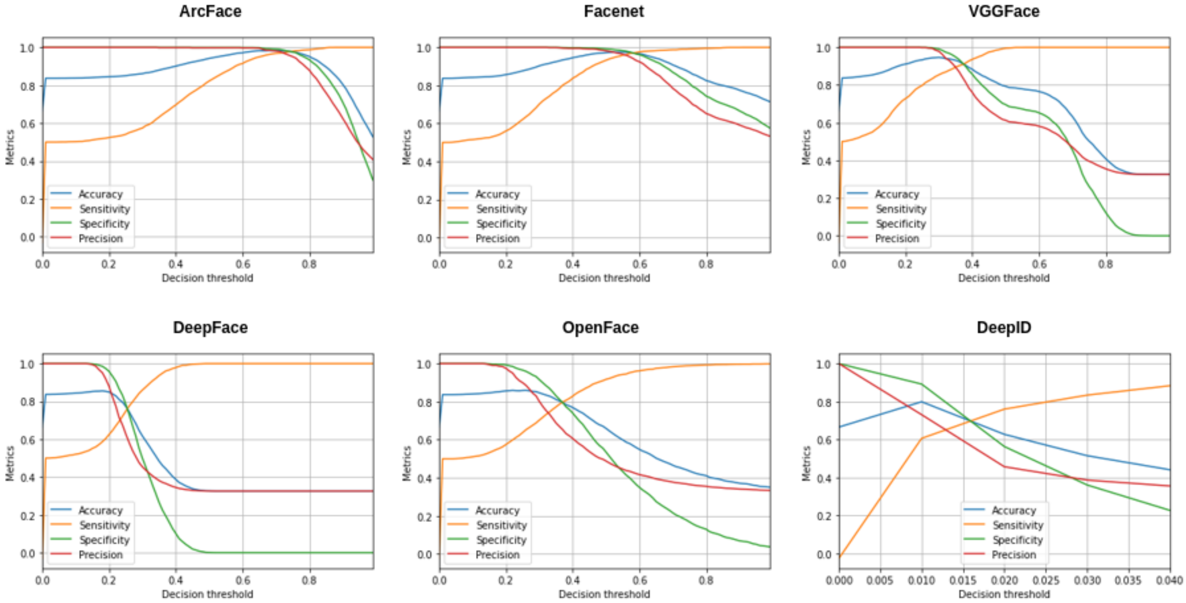

Fig. 3. Performance Graphs for Each Face Recognition Model

Table 2

Comparison of the face recognition models

Model	Parameters	Size, MB	MAC	FLOPS	Accuracy	Precision	Sensitivity	Specificity
VGGFace	145,002,878	553.2	15.5 bn	31.0 bn	0.78	0.63	0.81	0.77
DeepFace	102,412,864	390.7	0.50 bn	1.00 bn	0.74	0.58	0.78	0.73
ArcFace	34,165,184	130.8	4.47 bn	8.90 bn	0.98	0.98	0.96	0.99
Facenet	22,808,144	88.2	1.40 bn	2.80 bn	0.97	0.99	0.92	0.99
OpenFace	3,743,280	14.7	0.20 bn	0.40 bn	0.78	0.63	0.81	0.77
DeepID	395,080	1.55	0.08 bn	0.16 bn	0.68	0.52	0.71	0.67

The aforementioned models vary significantly in size and complexity, from the very large VGGFace with over 145 million parameters to the small DeepID model with only 395,000 parameters. In general, large models such as VGGFace and DeepFace have higher computational requirements: VGGFace requires approximately 15.5 billion MACs and 31 billion FLOPS. Smaller models, such as OpenFace and DeepID, are much less computationally intensive, requiring only 0.2-0.08 billion MACs and 0.4-0.16 billion

FLOPS. Accuracy is generally higher for larger models: VGGFace, DeepFace, ArcFace, and FaceNet achieve accuracy from 0.74 to 0.98. The smallest model, DeepID, has lower accuracy at 0.68. Metrics for precision, sensitivity, and specificity show similar trends. The highest accuracy of 0.98 was achieved by the ArcFace model, which maintains high performance with lower computational requirements compared to VGGFace and DeepFace. FaceNet achieves similar accuracy to ArcFace with even fewer parameters and lower MAC/FLOPS requirements.

Table 3

Comparison of the face recognition models in multi-class classification performance

Model	True Positives	False Negatives	Trur Negatives	False Positives
VGGFace	0.78	0.22	0.88	0.12
DeepFace	0.61	0.39	0.61	0.39
ArcFace	0.92	0.08	0.91	0.09
Facenet	0.89	0.11	0.94	0.06
OpenFace	0.68	0.32	0.70	0.30
DeepID	0.47	0.53	0.60	0.40

The models exhibit varying levels of ability to correctly identify the true positive (TP) compared to false negative (FN) identification of the correct person. They also differ in their ability to correctly recognize when images do not match (true negatives - TN) compared to false positive (FP) identification of an incorrect person. The ArcFace model shows the best results with a TP rate of 0.92 and a TN rate of 0.91. This means it correctly identifies the correct person 92% of the time and also correctly identifies mismatched images 91% of the time. Its false positive rate of 0.09 is also among the lowest. FaceNet also performs well with a TP rate of 0.89 and a high TN rate of 0.94, meaning it rarely incorrectly matches images to the wrong person. VGGFace, DeepFace, and OpenFace achieve more modest TP rates of 0.61-0.78, with higher FN and FP rates. The DeepID model has the most difficulty, with a TP rate of 0.47, indicating that less than half of its attempts are correct. Its TN rate of 0.60 also points to a high false positive rate of 0.40. In summary, ArcFace and FaceNet stand out with their high face recognition capabilities based on the balance of correct identifications (TP), correct rejection of mismatches (TN), and minimization of false negatives and false positives.

Conclusion

This work has investigated and compared several modern algorithms and models for face detection and recognition tasks in images. Comparisons were made based on key performance metrics such as accuracy, speed, and computational complexity. The analysis of face detection algorithms showed the advantages of the latest deep learning approaches, particularly the RetinaFace, DSFD, and MTCNN models. These models demonstrated high detection accuracy for faces of various sizes and in challenging conditions. However, they have the disadvantage of relatively low speed, especially for the most accurate DSFD (47.74 seconds). On the other hand, lighter models like SSD and DLib-HOG proved to be the fastest (0.055 seconds and 0.155 seconds, respectively), though their accuracy is somewhat lower. Overall, there is a trade-off between the accuracy and speed of detectors.For the task of recognizing detected faces, the best results were demonstrated by the latest models ArcFace and FaceNet. They combined high recognition accuracy of around 0.98 with relatively low computational requirements compared to the largest models like VGGFace. Thus, for creating optimal computer vision systems for faces, it is necessary to carefully select algorithms and models considering the requirements for accuracy, speed, and computational resources for the specific application. RetinaFace models for detection and ArcFace/FaceNet for recognition can be a good option to ensure a balance of these metrics. Overall, the research results demonstrate significant achievements in the field of computer vision and image processing thanks to deep learning methods. Further development of these technologies will allow for the creation of increasingly effective and versatile face detection and recognition systems.

References

[1] Y. Feng, S. Yu, H. Peng, Y. -R. Li and J. Zhang, "Detect Faces Efficiently: A Survey and Evaluations," in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 4, no. 1, pp. 1-18, Jan. 2022, doi: 10.1109/TBIOM.2021.3120412.

[2] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[3] B. Meden et al., "Privacy–Enhancing Face Biometrics: A Comprehensive Survey," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4147-4183, 2021, doi: 10.1109/TIFS.2021.3096024.

[4] T. Bezdan, N. Bačanin Džakula, "Convolutional Neural Network Layers and Architectures," in Sinteza 2019 - International Scientific Conference on Information Technology and Data Related Research, Belgrade, Singidunum University, Serbia, 2019, pp. 445-451. doi:10.15308/Sinteza-2019-445-451.

[5] Yi-Qing Wang, An Analysis of the Viola-Jones Face Detection Algorithm, Image Processing On Line, 4 (2014), pp. 128–148. doi:10.5201/ipol.2014.104.

[6] RetinaFace: Single-stage Dense Face Localisation in the Wild, Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, Stefanos Zafeiriou, 2019. doi:10.48550/arXiv.1905.00641.

[7] Face Image Feature Extraction based on Deep Learning Algorithm, Qing Kuang, 2021. doi:10.1088/1742-6596/1852/3/032040.

*[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Lecture Notes in Computer Science*, Springer International Publishing, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.*

[9] N. Zhang, J. Luo and W. Gao, "Research on Face Detection Technology Based on MTCNN," 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 2020, pp. 154-158, doi: 10.1109/ICCNEA50255.2020.00040.

[10]J. Li et al., "DSFD: Dual Shot Face Detector," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5055-5064, doi: 10.1109/CVPR.2019.00520.

*[11]M. K. Hasan, M. S. Ahsan, S. H. S. Newaz, and G. M. Lee, "Human face detection techniques: A comprehensive review and future research directions," *Electronics*, vol. 10, no. 19, p. 2354, 2021, doi: 10.3390/electronics10192354.*

[12]B. Dey, K. Khalil, A. Kumar and M. Bayoumi, "A Reversible-Logic based Architecture for VGGNet," 2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), Dubai, United Arab Emirates, 2021, pp. 1-4, doi: 10.1109/ICECS53924.2021.9665605.

[13]F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 815-823, doi: 10.1109/CVPR.2015.7298682.

*[14]B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," in *Proceedings of the 2016 Conference on Vision and Pattern Recognition* (CVPR), 2016.*

[15]Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220.

[16]W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, and X. Tang, "DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection," arXiv preprint arXiv:1412.5661, 2015.

[17]ArcFace: Additive Angular Margin Loss for Deep Face Recognition Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, Stefanos Zafeiriou, 2018., doi: 10.48550/arXiv.1801.07698.

ДОСЛІДЖЕННЯ СУЧАСНИХ МОДЕЛЕЙ НА ОСНОВІ ГЛИБОКОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ТА РОЗПІЗНАВАННЯ ОБЛИЧЧЯ

А. Сидор[1], Д. Балаж, Ю. Вітровий[2], О. Капшій, О. Карпін[3], Т. Максимюк[2]

[1]ТЗОВ «ЄВРОСОФТВЕР-ЮЕЙ», 79008, Україна, Львівська обл., місто Львів, вулиця Личаківська, будинок, 33а
[2] Національний університет «Львівська політехніка» вул. С. Бандери, 12, 79013, Львів, Україна
[3]Infineon Technologies, вул. Луганська, 20, 79034, Львів, Україна

Проблема побудови системи розпізнавання обличчя стикається з численними викликами, такими як зміни освітлення, пози і вирази обличчя. Основні етапи цього процесу включають виявлення, вирівнювання, виділення ознак та представлення обличчя. Кожен з цих етапів має критичне значення для досягнення точної ідентифікації. У статті аналізуються та порівнюються сучасні алгоритми та моделі для виявлення та розпізнавання облич за їх здатністю правильно ідентифікувати справжні позитивні (TP) та справжні негативні (TN) випадки, мінімізуючи при цьому хибні негативні (FN) та хибні позитивні (FP) випадки у розпізнаванні облич. Класичні алгоритми та прості моделі, такі як MediaPipe, забезпечують найвищу швидкодію, але за рахунок меншої точності. Навпаки, складніші моделі, такі як RetinaFace, забезпечують більшу точність за рахунок зниження швидкодії. Для систем, які пріоритезують максимальну точність виявлення і мінімізацію пропущених облич, рекомендуються моделі, такі як DSFD або RetinaFace-Resnet50, незважаючи на їх повільну роботу та непридатність для реального часу. Якщо основною метою є максимальна швидкість виявлення і прийнятне пропускання облич у неконтрольованих умовах, тоді слід обрати рішення SSD для розпізнавання облич. Для додатків, що вимагають балансу між швидкістю та точністю, оптимальною є модель RetinaFace-MobilenetV1, яка забезпечує швидкість виявлення в реальному часі та задовільну точність. Модель ArcFace демонструє найкращі результати з показником TP - 0.92 та TN - 0.91, що вказує на високу точність як у визначенні правильної особи, так і у відхиленні невідповідних зображень. ArcFace також підтримує низький рівень FP - 0.09. FaceNet слідує з показником TP - 0.89 та вражаючим TN - 0.94, демонструючи свою здатність уникати неправильних співпадінь. На відміну від цього, VGGFace, DeepFace та OpenFace показують помірні показники TP між 0.61 та 0.78, у поєднанні з вищими рівнями FN та FP. Модель DeepID демонструє найнижчу продуктивність з показником TP - 0.47 та TN - 0.60, що відображає значні труднощі у точному розпізнаванні. Висновки підкреслюють важливість вибору моделей на основі точності, швидкості та ресурсних вимог, пропонуючи RetinaFace та ArcFace/FaceNet як хороші варіанти компромісу.

Ключові слова: *детектування облич, розпізнавання облич, згорткові нейронні мережі, виділення ознак.*