Journal of Applied Science and Education, 2024,
Vol. 04, Iss. 01, S. No. 069, pp. 1-8
ISSN (Online): 2583-1372

Face Detection and Counting: Recent Advances and Future Research Directions

Shyam Sundar Singh[1], Vineet Singh[2], Shikha Singh[3], Bramah Hazela[4]

[1,2,3,4]Department of Computer Science and Engineering, Amity University Uttar Pradesh, Lucknow, India
[1]Shyamsundarsingh0056@gmail.com, [2]vsingh@lko.amity.edu, [3]ssingh8@lko.amity.edu, [4]bhazela@lko.amity.edu

How to cite this paper: S. S. Singh, V. Singh, S. Singh, B. Hazela, "Face Detection and Counting: Recent Advances and Future Research Directions," *Journal of Applied Science and Education (JASE)*, Vol. 04, Iss. 01, S. No. 069, pp 1-8, 2024.

https://doi.org/10.54060/a2zjournals.jase.69

Received: 15/01/2024
Accepted: 25/03/2024
Online First: 25/04/2024
Published: 25/04/2024

Abstract

In recent years, face detection and counting have undergone a remarkable transformation due to the emergence of deep learning techniques, particularly Convolutional Neural Networks (CNNs). These methods have surpassed traditional computer vision approaches, benefiting from vast datasets and robust computational resources. CNN's, the cornerstone of contemporary face recognition systems, excel in accurately identifying faces across diverse conditions and environments. The ascendancy of CNNs in face detection stems from their capacity to autonomously learn hierarchical features from raw pixel data, obviating the need for manual feature engineering. Consequently, they yield resilient systems capable of accommodating real-world variations like lighting, pose, expression, and occlusion. Their scalability enables the efficient processing of extensive image databases, rendering them invaluable for applications requiring face counting amidst crowded scenes. Nonetheless, challenges persist, particularly in deploying these systems on resources-constrained devices. The computational complexity and storage demand of deep CNNs necessitate ongoing exploration of lightweight network architectures that balances accuracy with reduced computational footprint. The historical Viola-jones algorithm, while foundational, has been eclipsed by the superior performance of CNNs. Deep learning's prowess lies in its adaptability to various tasks and its capacity to continuously refine itself with more data. By learning features directly from data, CNNs excel in capturing intricate patterns crucial for precise face detection and counting. Deep learning, especially through CNNs, has revolutionized face detection and counting by delivering unparalleled accuracy and robustness. However, ongoing efforts are crucial to address challenges in the efficiency and accessibility of these systems for broader deployment across various applications and devices.

Keywords

Face detection , Face Counting, Viola- Jones algorithm, deep learning, CNN

1. Introduction

The goal of people counting systems is to automatically estimate the population in public or enclosed spaces [1]. The objective is to develop methods and systems for accurately and efficiently recognizing and counting the number of human faces in images or video frames. These systems find application in various fields such as malls, traffic management, and auditoriums, where large crowds are present. The aim is to provide an automated system that detects and counts people, ensuring accuracy while maintaining speed and efficiency. Face Detection is a branch of computer science that identifies the size and placement of human faces in random photographs. It recognizes and detects characteristics that resemble the human face while ignoring everything else, such as the body, trees, automobiles, and buildings. Understanding Human face recognition is an important and growing research subject in computer vision [2]. Machines can distinguish things in photos and movies thanks to computer vision. Machines with CV skills identify and classify many items faster than humans. Human face identification and verification are major study fields in the computer vision field, and it has various applications, such as security cameras in malls and streets [3]. Face

Detection determines if an image contains faces or not and if faces are discovered in the picture, the position with a bound box for each face in the image is returned. Face detection is a fundamental component of face verification in a variety of applications [4]. Deep neural network (DNN) advancements pose the challenge of massive demand for data annotations. However, collecting item-level bounding box annotations, which are often required for training DNN-based object detection algorithms, is both expensive and time-consuming, especially for photos containing hundreds of objects [5].

The organization of the paper is as follows. It starts with an introduction in section 1 followed by an exhaustive literature survey in section 2. The methodology adopted is discussed in section 3, thereafter in section 4 current challenges and future directions has been talked about. Performance evaluation of the algorithms is explored in section 5 and the paper is ended through concluded in section 6.

2. Literature survey

2.1. History Perspective

The evolution of face identification spans from early rule-based methods to deep learning. Eigen faces and Viola-Jones pioneered in the 1980s and 2001, respectively, followed by SVMs. The introduction of CNNs in 2012 revolutionized real-time applications, accelerated by models like YOLO and SSD. Ongoing efforts focus on deep learning advancements, addressing posture variations, and dataset expansion, promising a bright future for face identification [6]. Fig.1 illustrates the significant shifts in image processing during the 2010s.

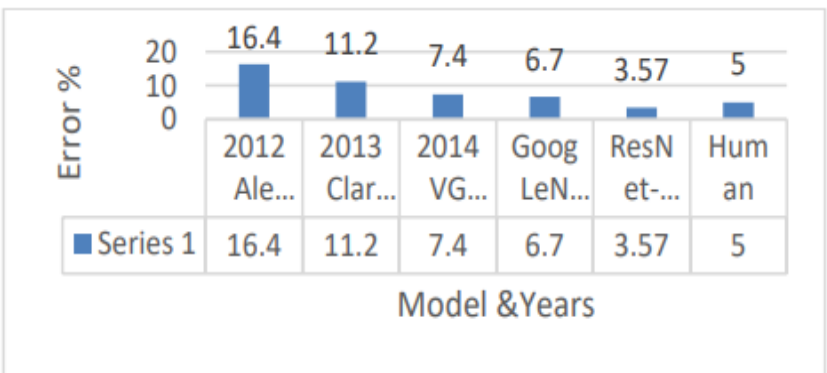

Figure 1. Error based years [6]

The acceptable categorization error rate in this period was approximately 25%. However, with the introduction of AlexNet in 2012, a deep convolutional neural network (CNN), this rate dramatically improved to 15.3%, marking a substantial milestone as it surpassed existing algorithms by more than 10.8%. Notably, AlexNet's performance secured it the winning title in the ILSVRC that year. Subsequent advancements in the field further refined accuracy: ZFNet achieved an error rate of 14.8% in 2013, GoogLeNet/Inception reduced it to 6.67% in 2014, and ResNet achieved an impressive 3.6% error rate in 2015. This progression underscores the rapid evolution and impact of deep learning in image processing [6].

2.2. Traditional Approaches

Traditional face identification methods progressed from manual features to rule-based and classic machine learning algorithms. Template matching compared predefined face templates with image sections. Eigen faces employed Principal Component Analysis for facial trait detection. Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) used texture and gradient details for recognition. Viola-Jones, in 2001, introduced Haar-like features and Ad boost-trained classifiers for real-time identification. Despite contributions, old methods struggled with scale, posture, lighting, and occlusions. Deep learning, especially convolutional neural networks, emerged to overcome these limitations, revolutionizing face identification for enhanced accuracy and robustness [7].

2.2.1. Viola Jones Algorithm

The Viola-Jones algorithm, created in 2001 by Paul Viola and Michael Jones, is a landmark face identification approach recognized for its efficiency. It uses Haar-like characteristics and integral pictures to perform fast computations. The technique distinguishes between face and non-facial areas using a cascade of Adaboost-trained classifiers. Its cascade structure enables rapid rejection of non-face regions, making it ideal for real-time applications such as video surveillance and facial recognition.

Despite being an early approach, Viola-Jones laid the groundwork for contemporary face detection techniques [8]. In Fig.2 the Viola-Jones method uses four Haar features: edge, linear, center, and diagonal [8].

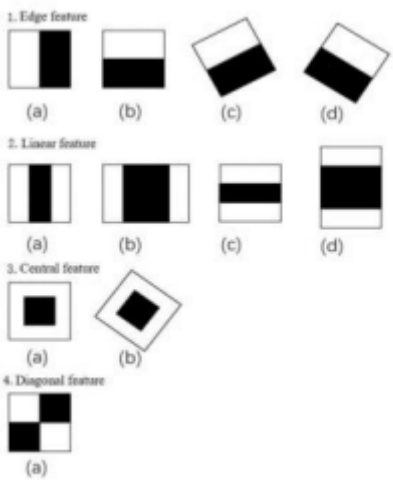

Figure 2. Four Haar Features [7]

2.2.2. Feature Based techniques and limitations

Feature-based face detection approaches, which are a fundamental strategy in the growth of facial recognition systems, entail identifying particular facial characteristics such as the eyes, nose, and mouth. However, these approaches have significant drawbacks. Their sensitivity to changes in position, lighting conditions, and facial emotions might compromise their perfor-

mance in real-world applications. Furthermore, manual feature engineering, a critical component of these strategies, necessitates domain knowledge and may lack adaptation to varied datasets. The inherent constraints of dealing with variances, as well as the requirement for user intervention, have driven a shift toward more complex and automated technologies, such as machine learning and deep learning, with the goal of improved face identification accuracy and resilience [9].

2.3. Machine Learning based approaches

Machine learning methods have been useful in improving facial detection algorithms. Face recognition technologies progressed from rule-based and handcrafted feature techniques to more data-driven and automated approaches as machine learning gained popularity. Support Vector Machines (SVMs) emerged as a key tool throughout this evolution. Faces were identified using SVMs, which learnt discriminatory patterns from attributes extracted directly from photographs [9]. Support Vector Machines (SVMs) excel in facial categorization, leveraging optimal decision boundaries from labeled datasets. Their flexibility and generalization prowess made them early favorites in face identification systems [10]. CNNs revolutionized face identification by automatically learning hierarchical features from raw pixel data, surpassing SVMs. This shift to deep learning eliminated laborious feature engineering, creating highly accurate and efficient face identification models.

3. Methodology

3.1. Deep Learning Technique

Deep learning, namely Convolutional Neural Networks (CNNs), has transformed face identification by automating feature extraction from raw visual data. CNNs extract hierarchical features, allowing for precise and efficient detection of face patterns. Object identification frameworks such as Faster R-CNN and SSD use region-based CNNs to improve bounding box prediction for faces. Real-time systems, such as You Only Look Once (YOLO), split pictures into grids to forecast bounding boxes and class probabilities. Innovations such as Multi-task Cascaded Convolutional Networks (lutional Networks (MTCNN) and Focal Loss in Retina Net demonstrate continued progress. These techniques improve accuracy, speed, and flexibility, making deep learning crucial in face identification for applications ranging from surveillance to human-computer interaction [10].

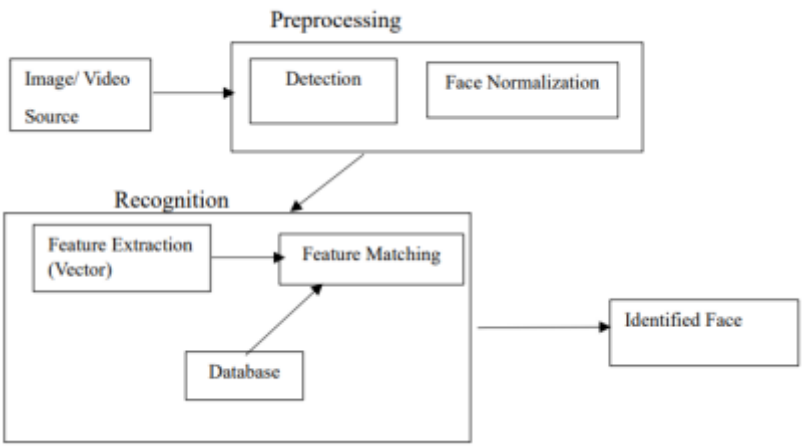

Figure3. Face recognition block diagram [10]

The face recognition block diagram in Fig. 3 illustrates the core components of a typical face recognition system. It typically includes stages such as face detection, feature extraction, and classification. Face detection identifies facial regions, feature extraction captures distinctive facial attributes, and classification matches them against known identities for recognition.

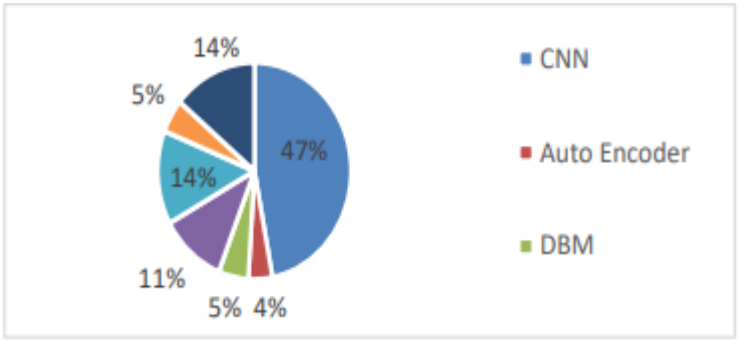

Figure 4. Different deep learning architectures for face recognition [10]

In Fig.4. different deep learning architectures for face recognition encompass CNN (Convolutional Neural Networks) for feature extraction, Autoencoders for unsupervised learning, DBM (Deep Boltzmann Machines) for probabilistic modeling, GANs (Generative Adversarial Networks) for generating realistic faces, Hybrid models combining various techniques, and Reinforcement Learning for adaptive recognition strategies, offering diverse approaches to address face recognition challenges [11] [12].

3.1.1. Convolutional Neural Network

The Convolutional Neural Network (CNN) is paramount in deep learning for image tasks like identification, classification, and feature extraction. It comprises feature extractors and classifiers. CNN operates through convolution, a linear operation between image matrices and kernels. An image matrix, representing pixels, has dimensions (HxWxD). Commonly, 3x3 kernels, such as edge detectors, are used. Here, H is the height, W is width, and D is the RGB color channel. Grayscale images have one channel, while color images have three (RGB). Kernels are matrices of size MxNxD. CNN's prowess lies in its ability to learn complex features hierarchically, making it indispensable in various image-related tasks. [13].

3.2. Face Counting Techniques

Face counting algorithms have emerged to solve the challenging challenge of determining the number of faces in pictures or video frames. This branch of computer vision has applications in a variety of domains, including crowd monitoring, surveil lance, and social behavior analysis. The strategies used in face counting are diverse, with each adapted to address distinct issues.

Density-based approaches use statistical techniques to determine the spatial distribution of faces in a picture. Kernel Density Estimation (KDE) and Gaussian Mixture Models (GMM) are examples of density-based approaches that use feature distribution to estimate face counts. These techniques are especially beneficial in congested environments when standard counting becomes difficult [5]. Regression-based techniques take a different approach, using machine learning models to directly estimate face count based on picture attributes. These models are trained using labeled datasets to understand the intricate link between visual attributes and face counts. Regression-based approaches are flexible and scalable, allowing for adaption to various datasets and settings [13].

Deep learning, particularly Convolutional Neural Networks (CNNs), has emerged as a dominant paradigm in face counting. CNNs use hierarchical characteristics to detect detailed patterns in congested surroundings. Deep learning models' capacity to automatically understand and represent complex relationships in data has considerably increased the accuracy of face counting systems. Researchers investigated alternative architectures and configurations to improve the performance of CNNs for face counting applications [8]. Detection and aggregation methods combine face detection models and counting

procedures. Individual faces in an image are identified by models such as Faster R-CNN or Single Shot MultiBox Detector (SSD), and their counts are combined to calculate the overall face count. This method is helpful in cases requiring exact face localization, such as security and surveillance applications [14].

Crowd counting datasets like ShanghaiTech and UCF_CC_50 aid in developing and evaluating face counting systems, enhancing their resilience and generalizability. Real-time video analysis enables continuous monitoring of face presence in crowded areas, crucial for scenarios like public events and smart city applications. However, challenges abound, including crowd density variations, occlusions, and the need for precise counting across diverse settings. Adaptation to dynamic real-world contexts, robustness against occlusions, and accounting for lighting conditions and demographic factors are critical for effective face counting systems [14].

4. Current Challenges and Future Directions

The field of face identification and counting has various modern issues that drive academics to continue exploring and innovating. One pressing worry is the detecting systems' capacity to adapt to a wide range of ambient variables, from illumination and weather fluctuations to complex background clutter. Furthermore, the complex challenge of controlling occlusions and overlapping faces in crowded settings necessitates advanced detection techniques. Achieving real-time speed without sacrificing precision is a constant challenge, especially in dynamic contexts. Furthermore, establishing the generality of face detection models across numerous populations, including various age groups, races, and gender presentations, is critical for equitable and impartial performance [15].

Researchers aim to advance face detection by strengthening adversarial robustness, integrating various modalities like optical and thermal data, and prioritizing explainability and ethical considerations. They emphasize privacy-preserving approaches such as federated learning and on-device processing while developing benchmark datasets to reduce biases and promote inclusion. Continuous learning methods enable systems to adapt dynamically, minimizing the need for frequent retraining.

Addressing security concerns, they design systems resilient to manipulative attacks. Emphasis lies on openness and responsibility in algorithmic decision-making, ensuring dependable and secure face detection systems capable of performing accurately even in adverse environmental conditions, thereby contributing to the broader societal well-being [16].

Human-centric evaluation measures are expected to play an important role in improving the assessment of face detection system performance. These KPIs will be more closely aligned with user experiences, taking into account the impact of false positives and false negatives on end users. Simultaneously, an investigation of the societal impact of broad face detection use would help academics get a thorough knowledge of topics such as surveillance, civil liberties, and unforeseen consequences [17].

5. Performance Evaluation Metrics

Performance evaluations should be quantitative. The report should show how many items were accurately recognized and how many false positives were generated. The assessment should allow one-to-one, one-to-many, and many-to-one matches. It should also be scalable to bigger test regions or numerous 3D scenarios without compromising tracking capacity [18].

The system shown in figure 7 and figure 8 includes three types of performance indicators: detection-based, tracking-based, and perimeter intrusion detection metrics. Detection-based metrics assess the performance of a System Under Test (SUT) on individual frames of video sensor data. They do not track the IDs of items during the exam. Each item is evaluated separately to ensure a match between the SUT and Ground-truth (GT) system for each video frame [19]. The experiment's performance score is calculated by averaging individual frame performance over all frames. Tracking-based metrics

compare GT and SUT tracks based on best correspondence, considering both the identification and entire trajectory of each item during the test run. The best matches are used to calculate error rates and performance indicators, as explained below.

The perimeter intrusion detection measure detects objects that enter a certain region [20].

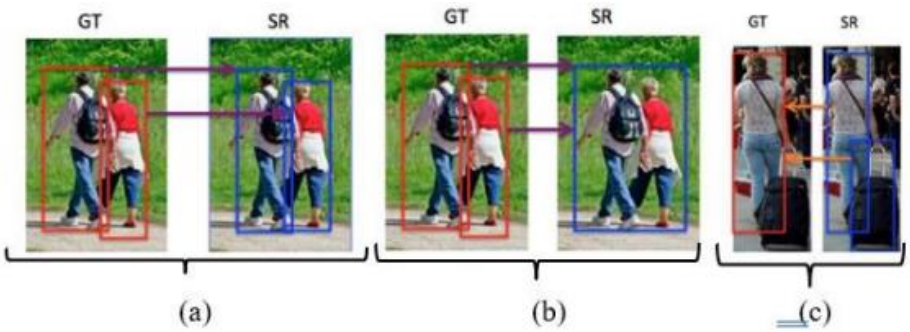

Fig.5. (a) One-to-one matching, (b) many-to-one matching and (c) one-to-many matching [18]

In Fig.5(a) One-to-one matching connects identified faces across frames, ensuring precise tracking and counting for enhanced monitoring accuracy. In Fig.5(b) Many-to-one matching links multiple faces in one frame to a single face in a previous frame, improving tracking and counting accuracy.

6. Conclusion and Future Scope

The landscape of face identification and counting poses challenges yet offers intriguing avenues for future exploration. Key findings stress adaptability to diverse environments, addressing occlusions and real-time performance while ensuring generalizability. Combining explainability and ethics is vital for transparent systems. Future research should enhance adversarial resilience, explore multi-modal fusion, and enable continuous learning for dynamic adjustments. Integrating privacy measures and diverse datasets mitigates biases and promotes inclusion. Human-centric assessment and societal impact analysis provide insight. Cooperation among researchers, industry, and lawmakers is crucial for ethical standards. Prioritizing human-centric evaluation connects technical advancements with social values. As progress continues, a comprehensive approach merging technology, ethics, and sociology will propel the field forward responsibly.

References

[1.] X. Zhao, E. Delleandrea and L. Chen, "A People Counting System Based on Face Detection and Tracking in a Video," *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, Genova, Italy*, pp. 67-72, 2009.

[2.] P. A. Mehta and T. J. Stonham, "A system for counting people in video images using neural networks to identify the background scene," *Journal of Pattern Recognition*, vol. 29, no. 8, pp. 1421–1428,

[3.] M. P. Tofiq Quadri, "Face Detection and Counting Algorithms Evaluation using OpenCV and JJIL," in *GITS-MTMIAt, Udaipur, Rajasthan, December*," vol. 2, p. 42, 2015.

[4.] M. S. Yehea Al Atrash, "Detecting and Counting People's Faces in Images Using Convolutional Neural Networks," *IEEE, no. Palestinian International Conference on Information and Communication Technology (PICICT)*, p. 116, 2021.

[5.] U. H., X. H. a. L.-P. C. Yi Wang, "A Self-Training Approach for Point-Supervised Object Detection and Counting in Crowds," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 30, no. 2876, p. 2876, 2021.

[6.] "https://anyconnect.com/blog/the-history-of-facial-recognition-technologies," [Online]

[7.] Y. S. a. H. C. Jing Huang, "Improved Viola-Jones face detection algorithm based on HoloLens," *EURASIP Journal on Image and Video Processing*, 2019.

[8.] L. W.-y. Ming, "Face Detection Based on Viola-Jones Algorithm Applying Composite Features," *International Conference on Robots & Intelligent System (ICRIS)*, p. 45, 2019.

[9.] A.F. D. S. MD. Tahmid Hasan Fuad, "Recent Advances in Deep Learning Techniques for Face Recognition," *IEEE Access,* p. 23, 2021.

[10.] P. W. S. C. H. Xudong Suna, "Face detection using deep learning: An improved faster RCNN approach," *Elsevier*, pp. 42-50, 2018.

[11.] J. J. L. C.-J. K. K. Sung Eun Choi, "Age face simulation using aging functions on global and local features with residual images," *Expert Systems with Applications*, pp. 80:107-125, 2017.

[12.] V. Rabaud and S. Belongie, "Counting crowded moving objects," *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pp. 705 – 711, 2006.

[13.] Mora Albiol and V. Naranjo, "Real-time high density people counter using morphological tools", *IEEE Trans. Intelligent Transportation Systems*, vol. 2, no. 4, pp. 204-218.

[14.] J. R.-d.-S. R. V. M. C. Gabriel Hermosilla, "A comparative study of thermal face recognition methods in unconstrained environments," *Pattern Recognition*, pp. 45(7): 2445-2459, 2012.

[15.] M. H. Mamta, "A new entropy function and a classifier for thermal face recognition," *Engineering Applications of Artificial Intelligence*, pp. 36: 269-286, 2014.

[16.] A.Y. Y. G. John Wright, *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 31, p. 457, 2, February 2009.

[17.] A. K. Jain, R. P. W. Duin and Jianchang Mao, "Statistical pattern recognition: a review," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[18.] R. B. S. Afzal Godil, "Performance Metrics for Evaluating Object and Human Detection and Tracking Systems," *Tracking Systems*, pp. 7972, 3-4.

[19.] J. Black, T. Ellis, and P. Rosin, "A Novel Method for Video Tracking Performance Evaluation", *The Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, October, Nice, France*, pp. 125-132.

[20.] J.C. Nascimento, J.S. and Marques, "Performance evaluation of object detection algorithms for video surveillance", *Multimedia, IEEE Transactions on* 8.4: 761-774, 2006.