


Article

Historical Blurry Video-Based Face Recognition

Lujun Zhai ¹ , Suxia Cui ¹, Yonghui Wang ^{2,*}, Song Wang ³, Jun Zhou ³ and Greg Wilsbacher ⁴

¹ Electrical and Computer Engineering Department, Prairie View A&M University, Prairie View, TX 77446, USA; lzhai@pvamu.edu (L.Z.); sucui@pvamu.edu (S.C.)

² Computer Science Department, Prairie View A&M University, Prairie View, TX 77446, USA

³ Computer Science and Engineering Department, University of South Carolina, Columbia, SC 29425, USA; songwang@cec.sc.edu (S.W.); zhouj@mailbox.sc.edu (J.Z.)

⁴ Moving Image Research Collections, University Libraries, University of South Carolina, Columbia, SC 29425, USA; gregw@mailbox.sc.edu

* Correspondence: yowang@pvamu.edu

Abstract: Face recognition is a widely used computer vision, which plays an increasingly important role in user authentication systems, security systems, and consumer electronics. The models for most current applications are based on high-definition digital cameras. In this paper, we focus on digital images derived from historical motion picture films. Historical motion picture films often have poorer resolution than modern digital imagery, making face detection a more challenging task. To approach this problem, we first propose a trunk–branch concatenated multi-task cascaded convolutional neural network (TB-MTCNN), which efficiently extracts facial features from blurry historical films by combining the trunk with branch networks and employing various sizes of kernels to enrich the multi-scale receptive field. Next, we build a deep neural network-integrated object-tracking algorithm to compensate for failed recognition over one or more video frames. The framework combines simple online and real-time tracking with deep data association (Deep SORT), and TB-MTCNN with the residual neural network (ResNet) model. Finally, a state-of-the-art image restoration method is employed to reduce the effect of noise and blurriness. The experimental results show that our proposed joint face recognition and tracking network can significantly reduce missed recognition in historical motion picture film frames.

Keywords: face detection; face recognition; face tracking; historical blurry video



Citation: Zhai, L.; Cui, S.; Wang, Y.; Wang, S.; Zhou, J.; Wilsbacher, G. Historical Blurry Video-Based Face Recognition. *J. Imaging* **2024**, *10*, 236. <https://doi.org/10.3390/jimaging10090236>

Academic Editor: Cosimo Distante

Received: 3 August 2024

Revised: 6 September 2024

Accepted: 14 September 2024

Published: 20 September 2024

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

From the 1890s through the 1970s, still and moving image celluloid film was the dominant method for documenting the world and its inhabitants. The vaults of government, corporate, and university libraries continue to house unknown hours of moving image film that contain the movements of historically important persons. Face recognition and object-tracking technology can help preserve the appearance of such persons [1,2].

Face recognition methods vary from traditional handcrafted machine learning to increasingly popular deep neural networks trained with big datasets [3]. Face recognition can be categorized into two classes: still image-based face recognition (SIFR) and video-based face recognition (VFR) [4]. Images in SIFR are typically captured by professional photographers under good conditions. Thus, sufficient facial features can be extracted by a neural network model to detect faces. When the video frame tends to be very motion-blurry and faces in videos exhibit occlusions and richer pose variations, VFR can be used to overcome these challenges [5–7].

Furthermore, historical videos have extrinsic features, noise, severe blurriness, etc., due to the low quality of video capture devices, which also exacerbates the lack of clarity in video frames [8,9]. These uncertain factors—noise, blurriness, lighting, occlusions, and pose variations—in historical video datasets can completely change the appearance of

figures [10–12] and severely impact facial feature extraction, significantly increasing the difficulty in accurately recognizing a particular face and causing failure of face recognition in consecutive frames, as shown in Figure 1. Thus, recognizing the faces in these videos and tracking a particular figure can be very challenging. Therefore, it is critical to have an effective face detector for applications involving blurry videos.

State-of-the-art face detection methods, such as you only look once (YOLO) [13], joint face detection and alignment using the face as the center point (CenterFace) [14], light and fast face detector (LFFD) [15], single-shot scale-invariant face detector (S3FD) [16], Two-Stage cascaded convolutional neural network (Two-Stage CNN) [17], attribute-aware face proposal and face detection network (Faceness-Net) [18], contextual multi-scale region-based convolutional neural network (CMS-RCNN) [19], dual-shot face detector (DSFD) [20], six degrees of freedom and face image pose estimation without face detection or landmark localization (img2pose) [21], and multi-task cascaded convolutional neural network (MTCNN) [22], have achieved impressive performance on standard still images or high-quality videos captured by current popular digital cameras. However, these models suffer from severe performance degradation on blurry and low-resolution historical videos. Therefore, a major difficulty of face recognition in historical videos is largely unsolved.

The goal of this study is to build a neural network model with an object-tracking algorithm for historical video-based face recognition. Specifically, we propose a trunk-branch concatenated multi-task cascaded convolutional neural network (TB-MTCNN) to enhance face detection performance. The trunk network learns facial features from blurry historical film images, while the branch network learns facial features from the low-level feature map generated by the trunk network. Richer facial features can be extracted with multi-scale receptive fields by incorporating various sizes of convolution filters into the network. Low-level layers are shared by the trunk and branch networks to reduce computational cost. To form the comprehensive facial features, the feature maps output by the trunk and branch networks are fused through concatenation.

Another aspect of our method is that the image face detection algorithm may fail in some frames; we can recover such missed faces by tracking the same object in continuous frames. Specifically, the TB-MTCNN model is applied as a face detector to extract facial features and predict the face bounding box. Deep SORT [23], an object-tracking method integrating the Kalman filter, the intersection-over-union (IOU) distance algorithm, and the Hungarian data association algorithm, is applied to face tracking and face bounding box updating. Subsequently, the updated detected face bounding box is transferred to the residual neural network (ResNet) [24] classifier to recognize a particular face; finally, we obtain the historical figure's face recognition. To better fit into face tracking, we optimize the object-based Deep SORT method utilizing the face-based Deep CNN model to efficiently extract facial features.

In addition, in order to reduce the effect of noise and blurriness and to fully extract facial features, we apply a state-of-the-art image restoration method, which brings old photos back to life [25] and is specifically designed for face restoration. The experimental results show that our proposed methods are promising in face detection in blurry videos and can effectively enhance face detection accuracy.



Figure 1. Face identification in historical motion picture video (President Johnson's face can be identified, as indicated by the red boxes, with prediction probabilities in frames 000433 and 000435, while face identification failed in frame 000434).

2. Related Works

2.1. Face Detection Using Deep Learning Algorithms

The concept of deep learning with neural networks was proposed many decades ago. Deep learning is a branch of machine learning methods based on artificial neural networks [26]. LeNet5 is the pioneering work in convolutional neural networks (CNNs), developed by researcher LeCun et al. in 1998 [27]. Due to the dramatic improvement in computing power and the explosion of big data, deep learning has made tremendous achievements over the past several years [28,29]. Problems related to sample scarcity can be solved by large datasets [30–32], and enhanced computing power has significantly accelerated the time-consuming training process. Deep learning-based approaches are increasingly applied in face detection fields and have significant advantages over traditional algorithms.

Region-based CNN methods have achieved state-of-the-art performance in generic object detection and classification [33], but satisfactory results have not been obtained in face detection. After training on a large-scale face detection dataset, the two-step approach Faster-RCNN [34], which employs a center loss function in the classification layer as an auxiliary signal, improved face detection performance to some extent. Faceness-Net [18] was proposed to address the issue of severe occlusions by leveraging facial attribute-based supervision. Img2pose [21] leverages the estimated 3D face pose to predict face bounding boxes. YOLO [13] and LFFD [15] mainly consider limited memory storage and low computing power and are designed for real-time edge device systems. Massively large receptive fields [35], as well as the EXT-D [36] and S3FD [16] models, aim to find extremely small faces. Two-Stage CNN [17] is a relatively simple and coarse multi-scale proposal generation-based face detector. CMS-RCNN [19] was proposed to detect relatively small and heavily occluded faces by integrating body context. The cascaded CNN network [37] was proposed for fast face detection. The first level of the network uses the dense sliding window for window sampling on the entire picture for classification but cannot achieve satisfactory results in small face detection. The optimized multi-task cascaded architecture with three stages of deep convolutional networks, MTCNN [22], combines face detection and alignment and predicts face and landmark locations simultaneously. MTCNN enables any size of image input into the network and uses the convolution operation to substitute the sliding window, which improves the efficiency of network operations. Few works [38–42] have achieved favorable performance on still images or blur-free videos.

As described above, face detection models were proposed mainly to reduce network running time, address the small face detection problem, or enhance performance on relatively clear image datasets captured by modern digital cameras with high resolution and less noise. However, historical video datasets, captured by low-quality devices due to immature camera technology, exhibit features with severe image blur and low resolution. The face detector models used in most common VFR applications cannot achieve the desired results on historical blurry videos. To learn facial features more effectively from blurry video, we propose a novel CNN-based model, TB-MTCNN. It consists of one trunk network and several branch networks. The trunk network learns facial features from raw face images, and each branch network learns representations from low-level feature maps generated by the trunk network. By concatenating the output feature maps from the trunk and branch networks, the model obtains the fused comprehensive facial features. Compared to MTCNN, TB-MTCNN is more accurate when evaluated on historical blurry videos and two publicly available large-scale face datasets: CASIA-WebFace [43] and Wider Face [17].

2.2. Face Recognition Algorithms

Face detection algorithms are able to detect the presence of a human face in an image or video, while face recognition goes further to recognize whose face it is. After faces are detected in historical videos using the proposed TB-MTCNN model, we need to employ an image recognition algorithm to further classify and identify the detected faces.

Deep convolutional neural networks [27,44] have led to a series of breakthroughs in image classification. VGGNet [45] improved classification accuracy by deepening the weight layers of AlexNet [44] to 16–19, using unified small convolution filters, and employing the rectified linear unit (ReLU) activation function. GoogleNet [46], a 22-layer deep network with multiple sizes of convolution filters was proposed to allow for increasing the depth and width of the network while considering the limitations of computational resources. ResNet [24] has been widely used in image recognition, as it adapts residual representation learning and shortcut connections and enables deeper CNNs to be easier to train and avoid the degradation of training accuracy. In this paper, we adapt ResNet [24] with 18 layers as the face classifier to recognize a specific person from the human faces detected by TB-MTCNN.

2.3. Face-Tracking Algorithms

Although face detection has been well developed in real application scenes, camera motion and the quality of video somewhat degrade face detection results and cause missed detection problems in consecutive video frames. Thus, researchers have devised face-tracking techniques that aim to effectively and robustly utilize appearance or motion information from previous frames to estimate the object's location in subsequent frames.

Many traditional methods like multiple hypothesis tracking (MHT) [47] and the joint probabilistic data association filter (JPDAF) [48] perform data association based on frame-by-frame information. Kim et al. proposed an appearance model for each target to prune the MHT graph [47]. In contrast, discriminative tracking methods like correlation filters (CFs) and kernelized correlation filters (KCFs) maintain a fixed target box throughout the tracking process [49]. However, changes in object sizes during the tracking process can lead to failures, as the correlation filters in these methods do not depend on detection information. Instead, they model the decision boundary and class differences. Predictive methods like the Kalman filter rely on detection information in the current frame and predict the most similar area in the next frame. SORT [50] and Deep SORT [23] employ the Kalman filter to anticipate facial positions in subsequent frames. Deep SORT enhances SORT by integrating appearance information, thereby improving tracking accuracy.

To efficiently account for missed facial recognition in consecutive video frames, we employ the Deep SORT algorithm as the face-tracking method in our proposed network.

2.4. Image Restoration

Old photo restoration is a classical mixed degradation problem. Most methods [51,52] focus on inpainting. Blurriness, noise, and blotches are identified based on low-level features and then inpainted using the textures nearby. Whyte proposed a non-uniform blind deblurring method based on a parametrized geometric model of the blurring process in terms of camera movement [53]. Noroozi employed a kernel-free end-to-end approach using a direct multi-scale CNN to deblur the images [54]. Image deblurring based on generative adversarial networks (GANs) [55–57] provided an effective tool to restore images and retrieve lost information. After deblurring, the recall rate of object detection increased on the synthetic datasets. However, these blind deblurring methods do not perform well on real-world images, as the blur is caused not only by the motion of different objects but also by camera jitter. To address this, a triplet-domain translation network was proposed [25], in which two variational autoencoders are trained to map old and clean photos into two separate latent spaces, effectively enhancing the visual quality of blurry photos to some extent. Ghosh explored the effects of nine kinds of degraded images and degradation removal on CNN-based image classification [58]. The degraded images decreased classification accuracy, while deblurring did not significantly improve classification accuracy.

Our work adapts one of the state-of-the-art image deblurring algorithms, which brings old photos back to life [25], to restore historical blurry videos and demonstrate the effectiveness of enhancing computer vision and face detection.

3. Face Recognition Structure

In this work, we make three main contributions to historical VFR. We propose a novel face detection neural network, TB-MTCNN, which efficiently extracts facial features from blurry historical videos by combining feature extraction from the trunk and branch networks. Further, to address the issue of missing detections, we integrate an object-tracking algorithm into the proposed face detection model. In addition, a state-of-the-art image restoration method is adopted to reduce the effects of noise and blurriness. The experimental results show that our proposed face recognition network significantly reduces missed detections in historical motion picture film frames, as detailed in Section 4.

Our neural network model consists of three parts: detection, classification, and tracking. The proposed TB-MTCNN network is based on the MTCNN algorithm [22], which performs the face detection and face alignment tasks simultaneously. We adopt ResNet18, a variant of the ResNet architecture introduced by He et al. [24], as a face classifier to recognize a specific person. In our neural network design, we integrate the Deep SORT algorithm into our model to track the face. It uses face detection information from previous frames to predict the target location in the current frame. In this way, the accuracy of face recognition is greatly increased. The architecture of the proposed model, integrating detection, recognition, and tracking, is shown in Figure 2.

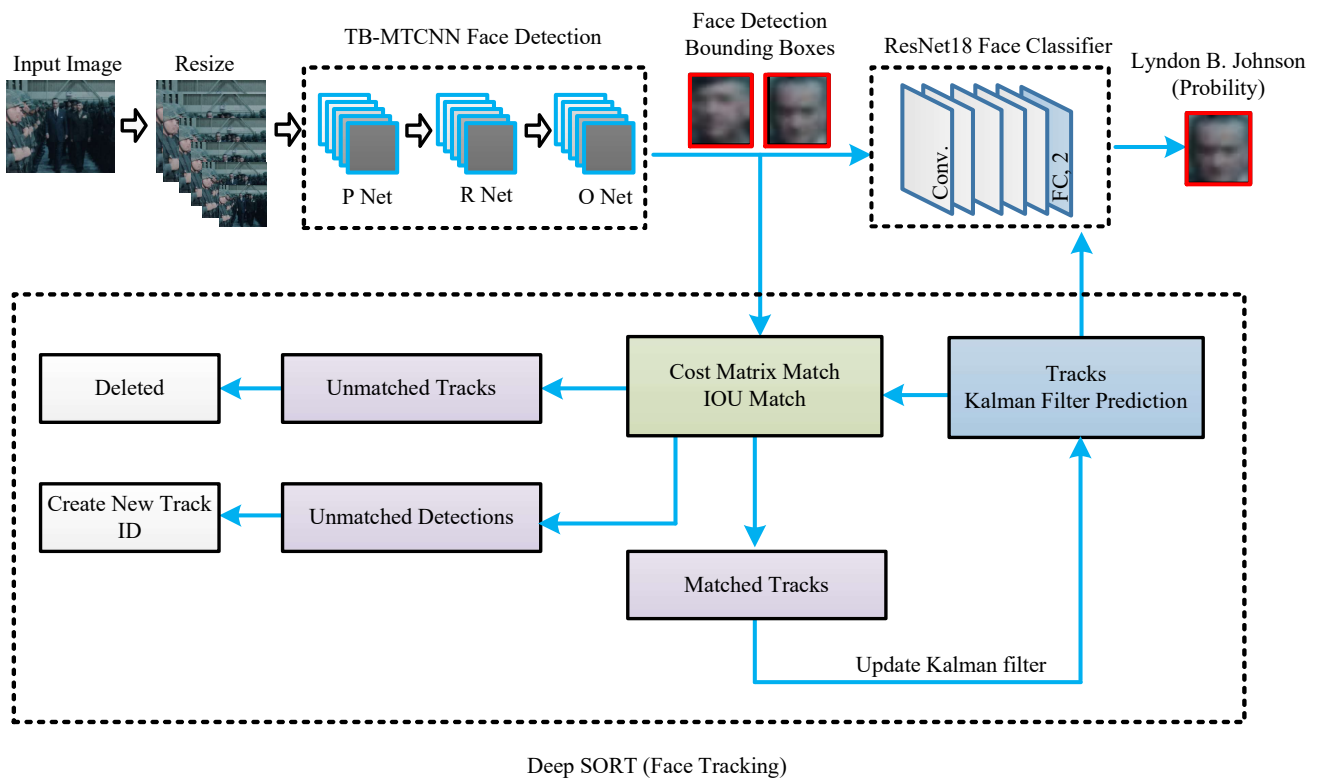


Figure 2. The architecture of the proposed model includes three components: a face detector (TB-MTCNN), a face tracker (Deep SORT), and a face classifier (ResNet18). First, TB-MTCNN processes the historical video frames to detect faces. Then, the Deep SORT algorithm utilizes face detection information from previous frames to track the face in the current frame if detection fails. Finally, the detection and tracking results are sent to the face classifier to determine whether the face belongs to President Johnson.

3.1. Face Detection

Face detection models used in common VFR applications cannot achieve the desired results in historical VFR, as video quality significantly affects face detection accuracy. To effectively extract rich facial features from historical videos, we propose a novel trunk–branch TB-MTCNN. The TB-MTCNN architecture is shown in Figure 3. The detailed

parameters of our proposed TB-MTCNN model are shown in Table 1. The TB-MTCNN model consists of three sub-networks: P-Net, R-Net, and O-Net. These three stages of deep convolutional networks predict face and landmark locations using a coarse-to-fine method. The purpose of each sub-network is described as follows:

- P-Net (Proposal Network): P-Net is the first stage in the TB-MTCNN cascade and plays a crucial role in generating preliminary face candidates across the image. It scans the image using a sliding window approach, allowing for the detection of faces at various locations.
 - Multi-Scale Detection: To capture faces of different sizes, P-Net operates on multiple scaled versions of the input image, effectively enabling the detection of both small and large faces within a single frame.
 - Probabilistic Assessment: Along with proposing candidate bounding boxes, P-Net also assigns a probability score to each box, indicating the likelihood that it contains a face. This score is critical for subsequent filtering processes in later stages.
 - Input Image Size: The network processes input images of size 12×12 pixels, optimized for rapid detection and low computational overhead, making it suitable for real-time applications.
- R-Net (Refinement Network):
 - Refinement of Proposals: After the initial detection by P-Net, R-Net refines the candidate bounding boxes by more accurately discerning which contain faces, significantly reducing the number of false positives.
 - Enhanced Spatial Resolution: R-Net processes higher resolution images (24×24 pixels) compared to P-Net's 12×12 pixels, enabling more effective discernment of facial features and improved localization of face boundaries.
 - Probability Filtering: R-Net computes a second-level confidence score to assess the likelihood of a face within each box, discarding boxes with low probability scores to reduce computational load and improve the precision of detections.
- O-Net (Output Network):
 - Final Bounding Box Adjustments: O-Net provides final adjustments to the bounding boxes, ensuring they are tightly fitted around the detected faces. This fine-tuning is crucial for applications requiring precise facial recognition or analysis.
 - Facial Landmark Detection: Beyond identifying faces, O-Net detects facial landmarks such as the eyes, nose, and mouth, supporting advanced facial analysis tasks like emotion recognition, facial alignment, and augmented reality applications.
 - High Receptive Field: With the largest input size (48×48 pixels), O-Net has a wider receptive field, allowing it to better integrate contextual information from a larger area of the image. This aids in accurate face and landmark localization even in complex scenarios.
 - Additional Outputs: The processing method of O-Net is similar to that of R-Net, but it also outputs face bounding box regressions and landmarks locations, in addition to performing face classification.

The trunk networks in each sub-network (colored in blue) take pyramid images as input and extract the coarse holistic facial features, while the branch networks (colored in brown) take the low-level feature maps generated by the trunk networks as input and further extract fine facial features. To extract facial features from multi-scale receptive fields, we integrate multiple sizes of convolution kernels into the branch networks, such as 1×1 , 3×3 , 5×5 , and 7×7 . By concatenating the output feature maps from the trunk and branch networks, the model obtains the fused, comprehensive facial features. Therefore, facial features can be extracted more completely.

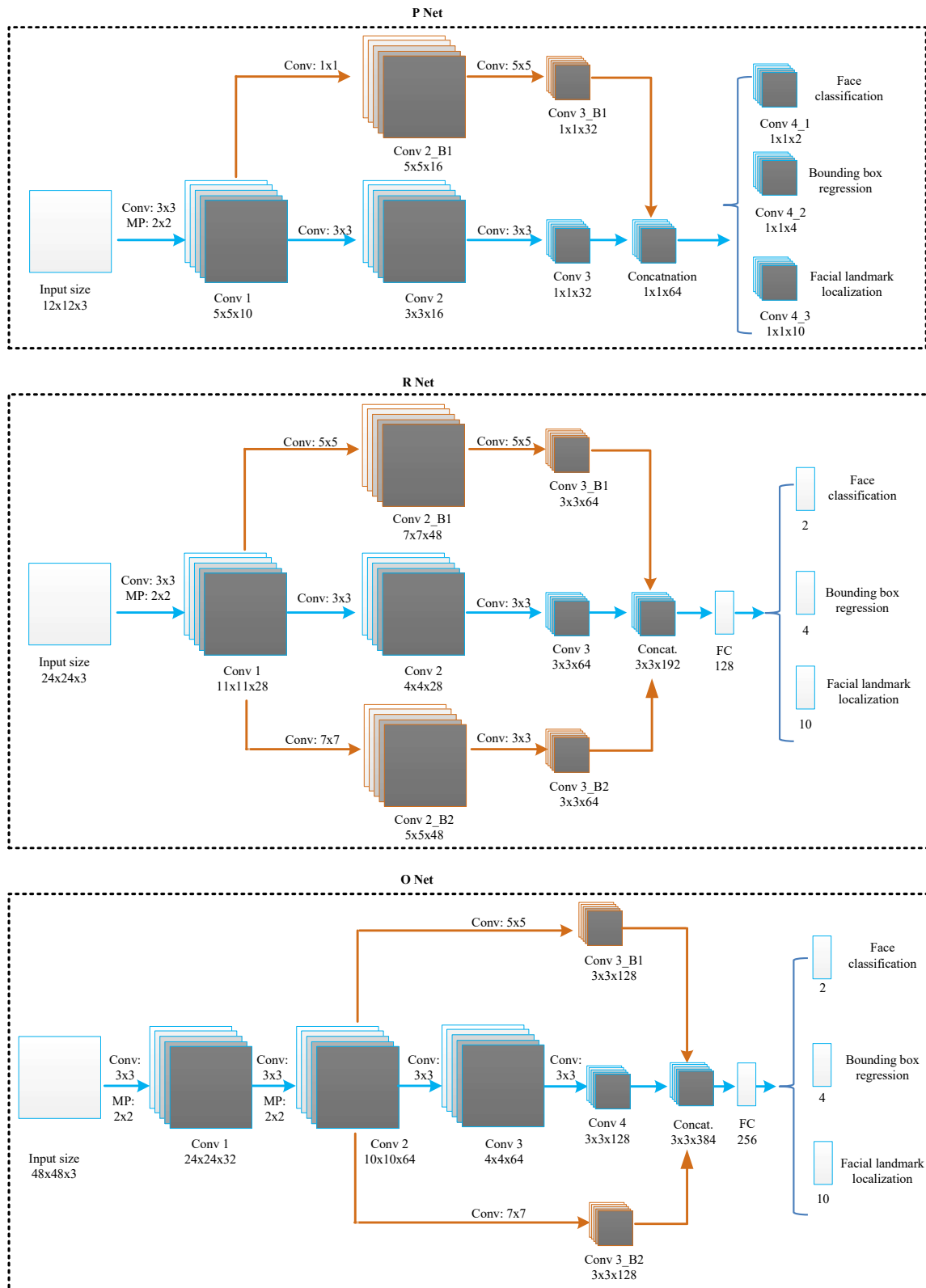


Figure 3. TB-MTCNN structure. The blue arrow represents the MTCNN architecture, while the yellow arrow highlights the integrated branch into the TB-MTCNN.

Table 1. Parameters in the TB-MTCNN model.

P-net	in shape	in channels	out channels	kernel size	stride	padding	out shape
<i>conv1</i>	[batch, 3, 12, 12]	3	10	3	1	0	[batch, 10, 10, 10]
<i>pool</i>	[batch, 10, 10, 10]	10	10	2	2	1	[batch, 10, 5, 5]
<i>conv2</i>	[batch, 10, 5, 5]	10	16	3	1	0	[batch, 16, 3, 3]
<i>conv2_B1</i>	[batch, 10, 5, 5]	10	16	1	1	0	[batch, 16, 5, 5]
<i>conv3</i>	[batch, 16, 3, 3]	16	32	3	1	0	[batch, 32, 1, 1]
<i>conv3_B1</i>	[batch, 16, 5, 5]	16	32	5	1	0	[batch, 32, 1, 1]
<i>conv4_1</i>	[batch, 32, 1, 1]	32	2	1	1	0	[batch, 1, 1, 1]
<i>conv4_2</i>	[batch, 32, 1, 1]	32	4	1	1	0	[batch, 4, 1, 1]
<i>conv4_3</i>	[batch, 32, 1, 1]	32	10	1	1	0	[batch, 10, 1, 1]
R-net	in shape	in channels	out channels	kernel size	stride	padding	out shape
<i>conv1</i>	[batch, 3, 24, 24]	3	28	3	1	0	[batch, 28, 22, 22]
<i>pool1</i>	[batch, 28, 22, 22]	28	28	3	2	1	[batch, 28, 11, 11]
<i>conv2</i>	[batch, 28, 11, 11]	28	48	3	1	0	[batch, 48, 9, 9]
<i>pool2</i>	[batch, 48, 9, 9]	48	48	3	2	0	[batch, 48, 4, 4]
<i>conv2_B1</i>	[batch, 28, 11, 11]	28	48	5	1	0	[batch, 48, 7, 7]
<i>conv2_B2</i>	[batch, 28, 11, 11]	28	48	7	1	0	[batch, 48, 5, 5]
<i>conv3</i>	[batch, 48, 4, 4]	48	64	2	1	0	[batch, 64, 3, 3]
<i>conv3_B1</i>	[batch, 48, 7, 7]	48	64	5	1	0	[batch, 64, 3, 3]
<i>conv3_B2</i>	[batch, 48, 5, 5]	48	64	3	1	0	[batch, 64, 3, 3]
line	in unit	out unit					
<i>line1</i>	192 × 3 × 3	128					
<i>line2_1</i>	128	1					
<i>line2_2</i>	128	4					
<i>line2_3</i>	128	10					
O-net	in shape	in channels	out channels	kernel size	stride	padding	out shape
<i>conv1</i>	[batch, 3, 48, 48]	3	32	3	1	0	[batch, 32, 46, 46]
<i>pool1</i>	[batch, 32, 46, 46]	32	32	2	2	1	[batch, 32, 24, 24]
<i>conv2</i>	[batch, 32, 24, 24]	32	64	3	1	0	[batch, 64, 22, 22]
<i>pool2</i>	[batch, 64, 22, 22]	64	64	3	2	0	[batch, 64, 10, 10]
<i>conv3</i>	[batch, 64, 10, 10]	64	64	2	1	0	[batch, 64, 8, 8]
<i>pool3</i>	[batch, 64, 8, 8]	64	64	2	2	0	[batch, 64, 4, 4]
<i>conv3_B1</i>	[batch, 128, 10, 10]	64	128	5	1	0	[batch, 128, 6, 6]
<i>pool3_B1</i>	[batch, 128, 6, 6]	128	128	2	2	0	[batch, 128, 3, 3]
<i>conv3_B2</i>	[batch, 64, 10, 10]	64	128	7	1	0	[batch, 128, 4, 4]
<i>pool3_B2</i>	[batch, 64, 4, 4]	128	128	2	1	0	[batch, 128, 3, 3]
<i>conv4</i>	[batch, 64, 4, 4]	64	128	2	1	0	[batch, 128, 3, 3]
line	in unit	out unit					
<i>line1</i>	384 × 3 × 3	256					
<i>line2_1</i>	256	1					
<i>line2_2</i>	256	4					
<i>line2_3</i>	256	10					

We leverage three tasks to train the TB-MTCNN model: face classification, bounding box regression, and facial landmark localization. The learning objective is a two-class classification problem. We use the cross-entropy loss:

$$L_i = -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)), \tag{1}$$

where L_i is the loss for the i -th training example, p_i is the probability predicted by the network that the i -th sample is a face, and y_i denotes the ground-truth label. For the bounding box regression problem, we employ the Euclidean loss:

$$L_i^{box} = \|\hat{y}_i^{box} - y_i^{box}\|_2^2, \tag{2}$$

where \hat{y}_i^{box} is the regression target predicted by our network and y_i^{box} represents the ground-truth coordinates.

Facial landmark detection is also a regression problem, and we use the Euclidean loss:

$$L_i^{landmark} = \|\hat{y}_i^{landmark} - y_i^{landmark}\|_2^2, \tag{3}$$

where $\hat{y}_i^{landmark}$ is the facial landmark's coordinates generated by the model and $y_i^{landmark}$ is the ground-truth coordinate value.

3.2. Face-Tracking Algorithm

To track faces in video frames, we adopt an object-tracking algorithm. This algorithm relies on face detection obtained from the current frame to predict face localization in the next frame. When the target fails to be detected by the TB-MTCNN model, the face-tracking algorithm can recover the missing face.

SORT (simple online and real-time tracking) [50] is a tracking method that employs Kalman filtering in image space and uses the Hungarian algorithm for optimal matching. Despite its efficiency, SORT struggles with occlusions, often assigning new track IDs to the same target when track continuity is temporarily lost. To enhance robustness against target losses, we incorporate the Deep SORT method. Deep SORT extends SORT by integrating a motion model with appearance information, allowing for more effective tracking of objects through occlusions and interactions. To enhance robustness against target losses, we incorporate the Deep SORT method. Deep SORT extends SORT by integrating a motion model with appearance information, allowing for more effective tracking of objects through occlusions and interactions. Deep SORT uses the cosine distance to weigh the appearance similarity between detections and tracks. To exclude the impossible condition, the Mahalanobis distance is used. Additionally, it adopts a matching cascade architecture to address the measurement-to-track associations, which increases robustness against target omissions and occlusions.

To accurately incorporate motion information, the squared Mahalanobis distance between predicted Kalman states and detections is used:

$$d_{i,j}^{(1)} = (d_j - p_i)^T S_i^{-1} (d_j - p_i), \tag{4}$$

where d_j represents the position of the j -th bounding box detection and p_i denotes the predicted position of the target from the i -th tracker. S_i is the covariance matrix that quantifies the uncertainty between the detection position and the predicted tracking position. $d_{i,j}^{(1)}$ quantifies the squared Mahalanobis distance for the i -th tracker and the j -th detection, measuring the statistical distance between the predicted state and the actual detection under the assumption of a Gaussian distribution. The pair (p_i, S_i) represents the projection of the i -th track's distribution into the measurement space. We set the threshold value of the Mahalanobis distance $t^{(1)}$ to 0.95 to exclude unlikely associations. $b_{i,j}^{(1)}$ indicates whether the association between the i -th tracker and the j -th detection is admissible:

$$b_{i,j}^{(1)} = [d^{(1)}(i, j) \leq t^{(1)}], \tag{5}$$

where $d_{i,j}^{(1)}$ denotes the squared Mahalanobis distance between the i -th tracker and the j -th detection, and $t^{(1)}$ represents the threshold value for the Mahalanobis distance to exclude unlikely associations between the tracker and detection.

To address scenarios where motion information may be less discriminative, such as after long-term occlusions, we adopt the cosine distance as a second metric. This is supported by the appearance descriptor r_j for each detection d_j , normalized to have a unit norm. We create a gallery $R_k = \{r_k^i\}_{k=1}^L$ for each track k , consisting of the last 100 associated appearance descriptors. The second metric $d_{i,j}^{(2)}$ measures the smallest cosine distance between the i -th tracker and the j -th detection in the appearance space:

$$d_{i,j}^{(2)} = \min\{1 - r_j^T r_k^i \mid r_k^i \in R^i\}. \tag{6}$$

This approach leverages both motion and appearance information to improve tracking accuracy and identity recovery in complex scenarios.

To further refine face tracking, we adopt Deep SORT [23], which utilizes a Deep CNN appearance descriptor initially trained on the pedestrian dataset MARS [59]. Given our specific requirements for face tracking, we retrain the model using a face-based dataset rather than a person-based dataset, ensuring that the extracted features are optimally suited for our application. The training dataset was sourced from historical video footage, providing a robust basis for feature extraction.

3.3. Face Classifier

To complete the face classification task with two classes—President Lyndon B. Johnson and unknown—a face classifier ResNet18 [24] is integrated into our face recognition network. ResNet offers a significant advantage in that the residual connection or shortcut connection is built across the residual blocks to avoid overfitting, gradient explosion, and vanishing gradient problems in image classification [60]. ResNet18 performs the face classification task as follows: first, ResNet18 takes the faces predicted by TB-MTCNN as input. Then, the convolutional layers extract facial features. Lastly, the fully connected layer predicts the probability of the face being Johnson's or unknown. In order to recognize the specific face, we fine-tuned the pre-trained ResNet model, which was trained on the labeled faces in the wild (LFW) [61] dataset and has extracted generic facial features. The reason we adapted fine-tuning is that our training dataset containing historical videos with President Johnson's face is small. It is best to train only the classifier layer with this small dataset. The lower layers of the network contain generic facial features, while the top layers contain specific facial features. To fine-tune ResNet [24], we froze all the weights except for the last fully connected (FC) layer. This final layer was replaced with random weights, and only this layer was trained with our dataset to extract the specific facial features.

4. Experiments

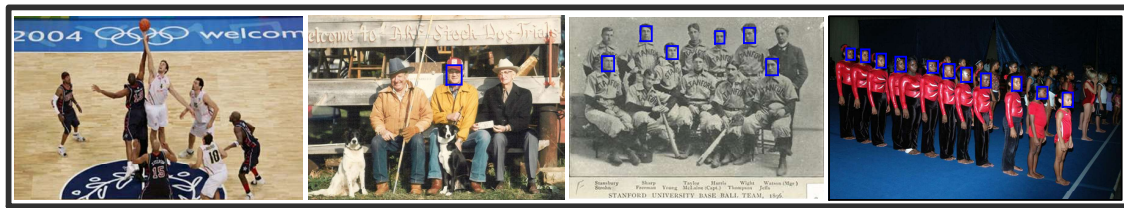
4.1. Effectiveness of the Proposed TB-MTCNN Model

Training Dataset: The proposed TB-MTCNN model was trained on a synthesized blurry images dataset. Image degradation was implemented on the publicly available Wider Face [17] dataset. This dataset includes 393,703 labeled faces with a high degree of variability in scales, poses, and occlusions, and is organized based on 61 different kinds of events. The training dataset was composed of negative, positive, partial, and landmark faces. Negative and positive samples were used for face classification tasks; positive and partial faces were used for bounding box regression tasks; landmark faces were used for facial landmark localization tasks. The image degradation included Gaussian blur and noise. The range for the random Gaussian blur kernel width was set to [1, 25]. The noise consisted of Gaussian noise as well as salt and pepper noise. The variance of Gaussian noise was specified within the range from 0 to 29, which corresponds to 0 to 29 percent of the maximum possible pixel value range, ensuring a relative measure across different image intensities. Salt and pepper noise levels were set within the range of [0, 0.005], quantifying the proportion of pixels affected.

Testing Dataset: To evaluate the proposed TB-MTCNN face detection model, we first conducted experiments on two publicly available large-scale face datasets: Wider Face [17] and CASIA-WebFace [43]. Then, we tested the model on a historical video dataset. The video dataset was obtained from the United States Marine Corps Film Repository. The online collection can be found at <https://digital.library.sc.edu/collections/united-states-marine-corps-films/> (accessed on 15 September 2020)

Performance Comparison on Wider Face [17]: This experiment evaluated the performance of the proposed TB-MTCNN model on the Wider Face Medium dataset. Figure 4 presents example images for the visual performance comparison between our approach and five state-of-the-art methods: Two-Stage CNN [17], Faceness [18], LFFD [15], CMS-RCNN [19], and img2pose [21]. Table 2 shows the quantitative results in terms of precision, recall, and F1 score. The proposed approach, TB-MTCNN, outperformed the original

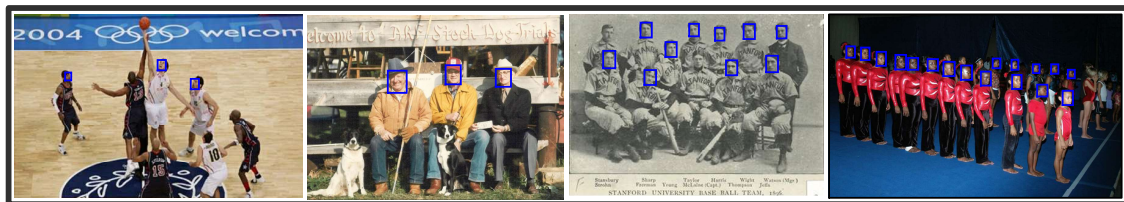
MTCNN model by as much as 7.3% in terms of the F1 score and achieved results comparable to the most advanced method, img2pose [21].



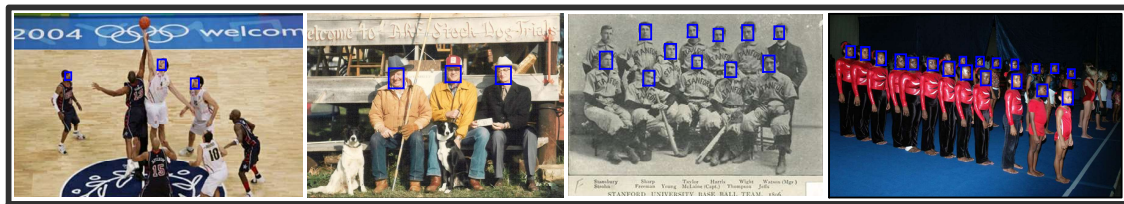
(a) Face detection using Two-stage CNN



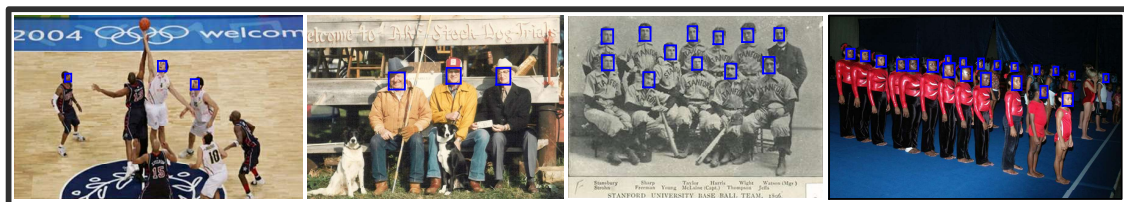
(b) Face detection using Faceness



(c) Face detection using LFFD



(d) Face detection using CMS-RCNN



(e) Face detection using img2pose



(f) Face detection using MTCNN



(g) Face detection using TB-MTCNN

Figure 4. Performance comparison of face detection models on the Wider Face Medium dataset.

Performance Comparison on CASIA-WebFace [43]: Next, we evaluated the performance of the face detection models on the CASIA-WebFace dataset, which contains 494,414 face images of 10,575 real identities collected from the web. Similar to the previous experiment, we applied these face detection models to the CASIA-WebFace testing dataset. Table 3 presents the performance comparison results. TB-MTCNN clearly outperformed the original MTCNN [22], Two-Stage CNN [17], and Faceness [18] in terms of the F1 score. The experimental results of TB-MTCNN, LFFD [15], CMS-RCNN [19], and img2pose [21] were very similar, but the performance of TB-MTCNN was slightly better than that of LFFD [15] and CMS-RCNN [19].

Table 2. Performance comparison of face detection models on the Wider Face Medium dataset.

Method	Precision	Recall	F1 Score
Two-Stage CNN [17]	0.589	0.496	0.539
Faceness [18]	0.604	0.617	0.610
LFFD [15]	0.865	0.693	0.769
CMS-RCNN [19]	0.874	0.704	0.779
img2pose [21]	0.891	0.735	0.805
MTCNN [22]	0.820	0.636	0.716
TB-MTCNN (present)	0.883	0.713	0.789

Table 3. Performance comparison of face detection models on the CASIA-WebFace dataset.

Method	Precision	Recall	F1 Score
Two-Stage CNN [17]	0.750	0.631	0.686
Faceness [18]	0.807	0.716	0.759
LFFD [15]	0.938	0.940	0.939
CMS-RCNN [19]	0.943	0.937	0.940
img2pose [21]	0.962	0.950	0.956
MTCNN [22]	0.925	0.892	0.908
TB-MTCNN (present)	0.964	0.942	0.945

Performance Comparison on Historical Videos: Finally, we compared the performance of the face detection models on historical videos. The United States Marine Corps Film Repository provides a visual record of Marines during World War I and World War II. The faces in the film exhibit a high degree of variability in scales, poses, and occlusions. Figure 5 and Table 4 present a visual comparison of the performance on sample images and the quantitative results, respectively. TB-MTCNN outperformed MTCNN [22], Two-Stage CNN [17], Faceness [18], LFFD [15], and CMS-RCNN [19] and exhibited similar performance to the most advanced model, img2pose [21].

Table 4. Performance comparison of face detection models on historical videos.

Method	Precision	Recall	F1 Score
Two-Stage CNN [17]	0.517	0.192	0.280
Faceness [18]	0.548	0.206	0.299
LFFD [15]	0.829	0.315	0.457
CMS-RCNN [19]	0.831	0.319	0.461
img2pose [21]	0.842	0.336	0.480
MTCNN [22]	0.750	0.223	0.344
TB-MTCNN (present)	0.836	0.325	0.468

Experimental Results Analysis: The proposed TB-MTCNN and five state-of-the-art methods exhibited consistent performance results across three different datasets. Two-Stage CNN, a coarse multi-scale proposal generation-based face detector, lacks a refined process for predicting faces; therefore its performance significantly lagged behind that

of TB-MTCNN and MTCNN [22]. Faceness [18] employs facial attribute supervision but lacks multi-scale facial feature extraction, leading to poorer performance in face detection. LFFD [15] and CMS-RCNN [19] achieved comparable results, but their performance was worse than that of TB-MTCNN. TB-MTCNN significantly outperformed the original MTCNN [22], as it combines the trunk and branch networks with multiple sizes of convolution kernels to enrich the receptive field and efficiently extract comprehensive facial features. Img2pose [21] slightly surpassed TB-MTCNN, as it adopts an indirect approach that leverages the estimated 3D face pose to achieve face detection. Compared to the relatively complex img2pose [21] algorithm, our proposed method is simple and can have a general application, making it easily integrable into all CNN-based face detection models. Specifically, branch networks with multiple sizes of convolution kernels, such as 1×1 , 3×3 , 5×5 , and 7×7 , can be introduced into CNN-based models. These branch networks take coarse feature maps from the low-level layers of the trunk network and further extract fine facial features from multi-scale receptive fields using these multiple sizes of convolution kernels. Finally, comprehensive facial features are formed by fusing the outputs of the branch and trunk networks. In this paper, the trunk network implementation is based on MTCNN [22], and the simplicity and effectiveness of the proposed method have been demonstrated.

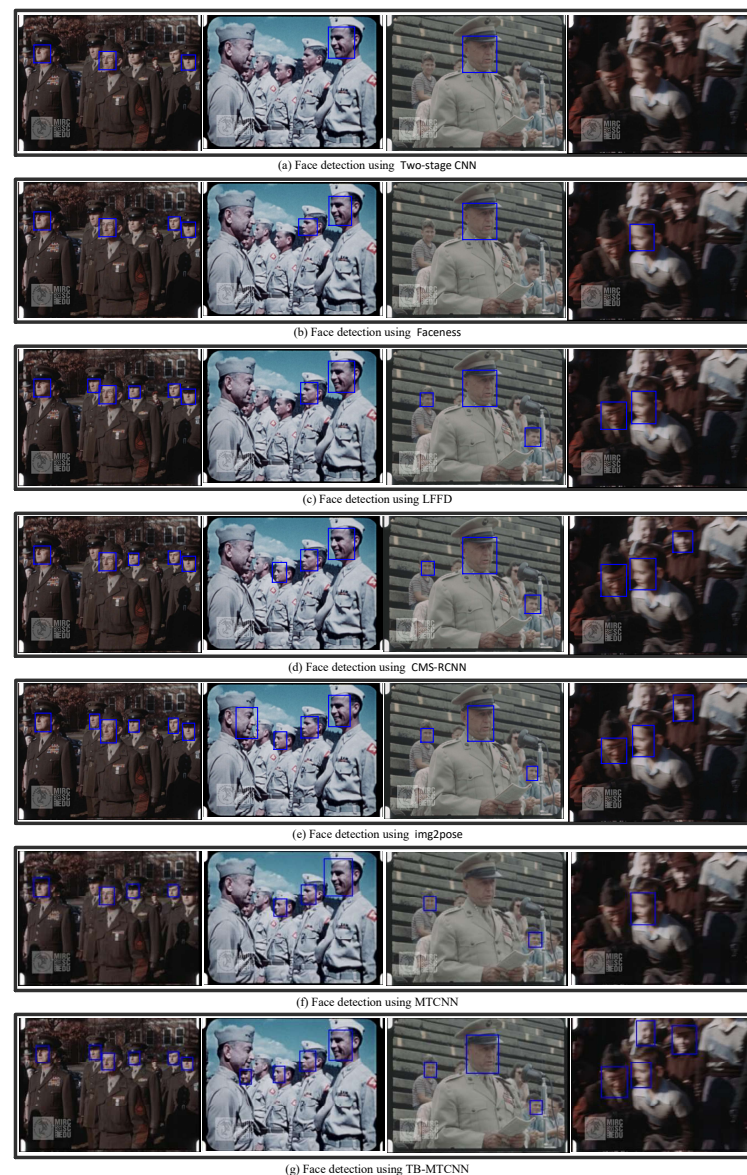


Figure 5. Sample performance comparison of face detection models on historical videos.

4.2. Fine-Tuning the Face Classifier

Collection of training and testing datasets: To train the face classifier ResNet18 [24] and recognize President Johnson's face, we collected images of President Lyndon B. Johnson and those of other people around him from historical films. The training dataset included 2000 images of President Johnson's face and 2000 images of unknown faces, while the testing data set included 500 images of President Johnson's face and 500 images of unknown faces. All images were resized to 320×320 .

Considering that our dataset contained noise and blurriness, to reduce these effects and extract facial features more effectively, we applied a state-of-the-art restoration method, which brings old photos back to life [25], to restore the images in both the training and testing sets. After restoration, we observed that the image quality improved from a visual perspective, as shown in Figure 6.

We conducted ResNet18 fine-tuning on the original blurry image dataset and the restored image dataset. The best training accuracy of the neural network model using the original data was 0.9312, while the best training accuracy using the restored images was 0.9528. This demonstrates that face-based image restoration can significantly improve face identification accuracy.



Figure 6. Effects of image restoration.

4.3. Training the Deep CNN as a Face Appearance Descriptor

To train the Deep CNN model, we annotated the ID and location information in the historical video film to create the training dataset. The dataset was split into two sets: 75% for training and 25% for testing. We set the total training duration to 100 epochs. The initial learning rate was set to 0.1, and we applied a decay factor that reduced the learning rate by multiplying it by 0.1 every 5 epochs. This frequent decay strategy was designed to ensure a more dynamic adjustment of the learning rate, promoting better convergence throughout training. The training accuracy on our evaluation dataset reached 0.9588, and the loss was 0.04.

4.4. Face Recognition and Tracking Experiments

Finally, we compared the historical VFR performance of TB-MTCNN with different approaches. The testing video film featured President Lyndon B. Johnson giving speeches and condolences to soldiers in the 1960s. We used the recall rate of President Johnson's face to evaluate the performance of each method. The results in Figure 7 clearly show that by combining the Deep SORT algorithm and deblurring images, the recall rate was 0.879, representing the best performance among the detection and tracking methods.

The experimental results indicate that when face recognition fails, the tracking algorithm can still accurately capture President Johnson's face position. For example, in Figure 7, we can see that using only the TB-MTCNN detector, President Johnson's face in frames 000015 and 000016 was recognized, as indicated by the red bounding boxes, but it failed in the subsequent two frames. After integrating the Deep SORT approach, President Johnson's face was tracked in frames 000017 and 000018, as indicated by the yellow bounding boxes. The object-tracking method successfully recovered the missing recognitions in these video frames. Table 5 shows the comparison results for various combinations of the proposed method.



Figure 7. The Deep SORT method recovered the missed detections (the red boxes are faces detected using the TB-MTCNN detector, and the yellow boxes are faces recovered using Deep SORT).

Table 5. Historical VFR performance comparison of different methods.

Method	Recall Rate	Precision
TB-MTCNN + ResNet	0.685	0.835
TB-MTCNN + ResNet + SORT	0.754	0.812
TB-MTCNN + ResNet + Deep SORT	0.863	0.834
TB-MTCNN + ResNet + Deep SORT+ Image Restoration	0.879	0.845

5. Conclusions

Compared with SIFR and common VFR, historical VFR is challenging in the field of facial recognition because of its unique features like severe blurriness, noise, and low resolution. In this paper, we proposed an efficient face detection model by combining trunk and branch networks and integrating various sizes of convolution kernels into the network to enrich the receptive field. It was demonstrated that our proposed trunk–branch concatenated neural network significantly improves face detection accuracy across several different image datasets. Furthermore, we proposed a joint face recognition and tracking network for historical video-based face identification, which enhances robustness against target omissions and recovers missing recognition in consecutive frames. Moreover, the employed image restoration method effectively reduces the impact of noise and blurriness on face recognition. The presented network’s effectiveness makes it well suited for face recognition in historical VFR.

Author Contributions: L.Z.: Led the technical aspects, including model development. S.C.: Secured funding for the project. Y.W.: Contributed with their research expertise, aligning the project with their primary research direction. S.W.: Provided detailed technical guidance, new insights, and experimental suggestions. J.Z.: Assisted with polishing and editing the manuscript. G.W.: Facilitated access to the historical video film repository. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by NSF grants OAC-1827243 and HRD-1411260, as well as the US Department of Education grant P120A180114.

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to privacy concerns related to the historical films from the Moving Image Research Collections, the data involved in this research will not be released

Acknowledgments: The authors would also like to thank the Extreme Science and Engineering Discovery Environment (XSEDE) Campus Champions Fellows Program, NSF award number ACI-1548562.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Carstensen, L.L. The Influence of a Sense of Time on Human Development. *Science* **2006**, *312*, 1913–1915. [[CrossRef](#)] [[PubMed](#)]
- Best, J.J. Who Talked to the President When? A Study of Lyndon B. Johnson. *Political Sci. Q.* **1988**, *103*, 531–545. [[CrossRef](#)]
- Sun, X.; Wu, P.; Hoi, S.C. Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing* **2018**, *299*, 42–50. [[CrossRef](#)]
- Ding, C.; Tao, D. Trunk-Branch Ensemble Convolutional Neural Networks for Video-Based Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1002–1014. [[CrossRef](#)]
- Hadid, A.; Pietikainen, M. From still image to video-based face recognition: An experimental analysis. In Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, Republic of Korea, 19 May 2004; pp. 813–818. [[CrossRef](#)]
- Li, Z.; Tie, Y.; Qi, L. Face Recognition in Real-world Internet Videos Based on Deep Learning. In Proceedings of the 2019 8th International Symposium on Next Generation Electronics (ISNE), Zhengzhou, China, 9–10 October 2019; pp. 1–3. [[CrossRef](#)]
- Huang, Z.; Shan, S.; Wang, R.; Zhang, H.; Lao, S.; Kuerban, A.; Chen, X. A Benchmark and Comparative Study of Video-Based Face Recognition on COX Face Database. *IEEE Trans. Image Process.* **2015**, *24*, 5967–5981. [[CrossRef](#)]
- Ong, E.P.; Loke, M.H.; Lin, W.; Lu, Z.; Yao, S. Video Quality Metrics—An Analysis for Low Bit Rate Videos. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing—ICASSP '07, Honolulu, HI, USA, 15–20 April 2007; Volume 1, pp. I-889–I-892. [[CrossRef](#)]
- Li, M.; Jianbin, S.; Hui, L. A Determining Method of Frame Rate and Resolution to Boost the Video Live QoE. In Proceedings of the 2nd International Conference on Multimedia and Image Processing (ICMIP), Wuhan, China, 17–19 March 2017; pp. 206–209. [[CrossRef](#)]
- Kharchevnikova, A.; Savchenko, A.V. Efficient video face recognition based on frame selection and quality assessment. *PeerJ Comput. Sci.* **2021**, *7*, e391. [[CrossRef](#)] [[PubMed](#)]
- Taskiran, M.; Kahraman, N.; Eroglu Erdem, C. Hybrid face recognition under adverse conditions using appearance-based and dynamic features of smile expression. *IET Biom.* **2021**, *10*, 99–115. [[CrossRef](#)]
- Handa, A.; Agarwal, R.; Kohli, N. Incremental approach for multi-modal face expression recognition system using deep neural networks. *Int. J. Comput. Vis. Robot.* **2021**, *11*, 1–20. [[CrossRef](#)]
- Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [[CrossRef](#)]
- Xu, Y.; Yan, W.; Sun, H.; Yang, G.; Luo, J. CenterFace: Joint Face Detection and Alignment Using Face as Point. *arXiv* **2019**, arxiv:1911.03599. [[CrossRef](#)]
- He, Y.; Xu, D.; Wu, L.; Jian, M.; Xiang, S.; Pan, C. LFFD: A Light and Fast Face Detector for Edge Devices. *arXiv* **2019**, arxiv:1904.10633.
- Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; Li, S.Z. S³FD: Single Shot Scale-invariant Face Detector. *arXiv* **2017**, arxiv:1708.05237.
- Yang, S.; Luo, P.; Loy, C.C.; Tang, X. WIDER FACE: A Face Detection Benchmark. *arXiv* **2015**, arxiv:1511.06523.
- Yang, S.; Luo, P.; Loy, C.C.; Tang, X. From facial parts responses to face detection: A deep learning approach. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3676–3684.
- Zhu, C.; Zheng, Y.; Luu, K.; Savvides, M. CMS-RCNN: Contextual Multi-Scale Region-based CNN for Unconstrained Face Detection. *arXiv* **2016**, arxiv:1606.05413.
- Li, J.; Wang, Y.; Wang, C.; Tai, Y.; Qian, J.; Yang, J.; Wang, C.; Li, J.; Huang, F. DSFD: Dual Shot Face Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5055–5064. [[CrossRef](#)]
- Albiero, V.; Chen, X.; Yin, X.; Pang, G.; Hassner, T. img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation. *arXiv* **2021**, arxiv:2012.07791.
- Chou, K.; Cheng, Y.; Chen, W.; Chen, Y. Multi-task Cascaded and Densely Connected Convolutional Networks Applied to Human Face Detection and Facial Expression Recognition System. In Proceedings of the 2019 International Automatic Control Conference (CACs), Keelung, Taiwan, 13–16 November 2019; pp. 1–6. [[CrossRef](#)]
- Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. *arXiv* **2017**, arxiv:1703.07402.
- He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arxiv:1512.03385.
- Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; Wen, F. Bringing Old Photos Back to Life. *arXiv* **2020**, arxiv:2004.09484.
- Høye, T.T.; Årje, J.; Bjerge, K.; Hansen, O.L.; Iosifidis, A.; Leese, F.; Mann, H.M.; Meissner, K.; Melvad, C.; Raitoharju, J. Deep learning and computer vision will transform entomology. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2002545117. [[CrossRef](#)]
- LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [[CrossRef](#)]
- Sadiku, M.N.O.; Zhou, Y.; Musa, S.M. Smart Computing. *Int. J. Eng. Res. Adv. Technol.* **2019**, *5*, 26–29. [[CrossRef](#)]
- Nandal, P.; Bura, D.; Singh, M. Emerging Trends of Big Data in Cloud Computing. In *Applications of Big Data in Large-and Small-Scale Systems*; IGI Global: Hershey, PA, USA, 2021; pp. 38–55.

30. Lim, H.i. A Study on Dropout Techniques to Reduce Overfitting in Deep Neural Networks. In *Advanced Multimedia and Ubiquitous Engineering: MUE-FutureTech*; Park, J.J., Loia, V., Pan, Y., Sung, Y., Eds.; Springer: Singapore, 2021; pp. 133–139.
31. Chen, T.; Zhang, Z.; Liu, S.; Chang, S.; Wang, Z. Robust overfitting may be mitigated by properly learned smoothing. In *Proceedings of the International Conference on Learning Representations*, Vienna, Austria, 4 May 2021; Volume 1.
32. Shi, X.; Liu, Y. Sample Contribution Pattern Based Big Data Mining Optimization Algorithms. *IEEE Access* **2021**, *9*, 32734–32746. [[CrossRef](#)]
33. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [[CrossRef](#)]
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
35. Hu, P.; Ramanan, D. Finding Tiny Faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 1522–1530. [[CrossRef](#)]
36. Yoo, Y.; Han, D.; Yun, S. EXT-D: Extremely Tiny Face Detector via Iterative Filter Reuse. *arXiv* **2019**, arxiv:1906.06579.
37. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 7–12 June 2015; pp. 5325–5334. [[CrossRef](#)]
38. Zhang, F.; Fan, X.; Ai, G.; Song, J.; Qin, Y.; Wu, J. Accurate Face Detection for High Performance. *arXiv* **2019**, arxiv:1905.01585.
39. Zhang, C.; Xu, X.; Tu, D. Face Detection Using Improved Faster RCNN. *arXiv* **2018**, arxiv:1802.02142.
40. Wang, Y.; Ji, X.; Zhou, Z.; Wang, H.; Li, Z. Detecting Faces Using Region-based Fully Convolutional Networks. *arXiv* **2017**, arxiv:1709.05256.
41. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708. [[CrossRef](#)]
42. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 7–12 June 2015; pp. 815–823. [[CrossRef](#)]
43. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning Face Representation from Scratch. *arXiv* **2014**, arxiv:1411.7923.
44. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
45. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
46. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arxiv:1409.4842.
47. Kim, C.; Li, F.; Ciptadi, A.; Rehg, J.M. Multiple Hypothesis Tracking Revisited. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015; pp. 4696–4704. [[CrossRef](#)]
48. Rezatofighi, S.H.; Milan, A.; Zhang, Z.; Shi, Q.; Dick, A.; Reid, I. Joint Probabilistic Data Association Revisited. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, hSantiago, Chile, 7–13 December 2015; pp. 3047–3055. [[CrossRef](#)]
49. Kim, C.; Fuxin, L.; Alotaibi, M.; Rehg, J.M. Discriminative Appearance Modeling with Multi-track Pooling for Real-time Multi-object Tracking. *arXiv* **2021**, arXiv:2101.12159.
50. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, 25–28 September 2016. [[CrossRef](#)]
51. Giakoumis, I.; Nikolaidis, N.; Pitas, I. Digital image processing techniques for the detection and removal of cracks in digitized paintings. *IEEE Trans. Image Process.* **2006**, *15*, 178–188. [[CrossRef](#)] [[PubMed](#)]
52. Chang, R.C.; Sie, Y.L.; Chou, S.M.; Shih, T. Photo Defect Detection for Image Inpainting. In *Proceedings of the 7th IEEE International Symposium on Multimedia (ISM'05)*, Irvine, CA, USA, 14 December 2005; p. 5. [[CrossRef](#)]
53. Whyte, O.; Sivic, J.; Zisserman, A.; Ponce, J. Non-uniform deblurring for shaken images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 13–18 June 2010; pp. 491–498. [[CrossRef](#)]
54. Noroozi, M.; Chandramouli, P.; Favaro, P. Motion Deblurring in the Wild. *arXiv* **2017**, arxiv:1701.01486.
55. Fan, S.; Luo, Y. Deblurring Processor for Motion-Blurred Faces Based on Generative Adversarial Networks. *arXiv* **2021**, arxiv:2103.02121.
56. Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; Matas, J. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. *arXiv* **2018**, arxiv:1711.07064.
57. Lenka, M.K.; Pandey, A.; Mittal, A. Blind Deblurring Using GANs. *arXiv* **2019**, arxiv:1907.11880.
58. Ghosh, S.S.; Hua, Y.; Mukherjee, S.S.; Robertson, N.M. Improving Detection And Recognition Of Degraded Faces By Discriminative Feature Restoration Using GAN. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, UAE, 25–28 October 2020; pp. 2146–2150. [[CrossRef](#)]
59. Wojke, N.; Bewley, A. Deep Cosine Metric Learning for Person Re-Identification. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 748–756. [[CrossRef](#)]

-
60. Koonce, B. ResNet 50. In *Convolutional Neural Networks with Swift for Tensorflow*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 63–72.
 61. Huang, G.B.; Ramesh, M.; Berg, T.; Learned-Miller, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*; Technical Report 07-49; University of Massachusetts: Amherst, MA, USA, 2007.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.