A Frame Work Designing for Deep Fake Motion Detection using Deep Learning in Video Surveillance Systems

Srikanth Bethu[1*], M. Ratna Sirisha[2], Kothai Andal C[3], Gayathri R[4], Chandramouli H[5], R. Aruna[6]

[1,2]Department of CSE, CVR College of Engineering, Hyderabad-501510, Telangana, India
[3]Department of EEE, AMC Engineering College, Bengaluru-560083, Karnataka, India
[4]Department of ECE, Sree Dattha Institute of Engineering and Science, Hyderabad-501510, Telangana, India
[5]Department of CSE, East Point College of Engineering and Technology, Bangalore-560049, Karnataka, India
[6]Department of ECE, AMC Engineering College, Bengaluru-560083, Karnataka, India

srikanthbethu@gmail.com,msiri515@gmail.com

Abstract. This exploration centers around constant perception of items in a given setting, which prompts a rundown of the ways of behaving or connections of the things. Because of the absence of time for robotized checking and examination, an administrator should watch a lot of video information in time with complete thoughtfulness regarding recognize any inconsistencies or episodes, or solely after the surprising occurrence has happened may the video information be utilized as proof. To overcome these issues, we have developed a Deep Learning model to detect face objects in real-time to identify their motion for fake detection using Video Surveillance systems. We have also compared our model with the existing models, and we can probably secure high accuracy of 95.4%.

Keywords: Face Motion, Deep Learning, Deep Fake detection.

1 Introduction

Video surveillance is a significant concern in computer vision research to identify, recognize, and track data across a series of images and to understand and explain object behavior by replacing the outdated old method of human operators operating cameras [1]. A computer vision system can detect immediate unauthorized activity as well as long-term suspicious activities are also possible, alerting a human operator to conduct a more thorough investigation. The control operator of a manual video surveillance system conducts all tasks while observing the visual information from the different cameras.

Face discovery and acknowledgment in video successions is a provoking issue because of the trouble of the capability of low-quality plans and scene intricacies. Video-

C. Kiran Mai et al. (eds.), *Proceedings of the Fourth International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023)*, Atlantis Highlights in Computer Sciences 18,
https://doi.org/10.2991/978-94-6463-314-6_18

based acknowledgment frameworks offer crucial data to upgrade the reconnaissance framework contrasted with single picture-based acknowledgment frameworks. Images can offer a great deal of data about an object, but they also present a number of difficulties. As a result, they are encouraging the development of a process for extracting key features from the entire casing.

Face location, checking, and acknowledgment are conceivable from video arrangements utilizing different procedures. In addition, due to the different lighting and poses, it is very difficult to identify the person facial image. Thus, computing and incorporating these elements across a few casings takes excellent work. However, as this method requires more memory and time, it becomes tedious to identify human faces in video sequences.

Motion detection is the initial stage of a video surveillance system. Motion detection separates the moving foreground element from the remainder of the image. In the video [2], effective segmentation of foreground objects aids subsequent processes such as object detection and behavior recognition. Context subtraction, optical flow, and temporal differencing are used to segment motion. Background subtraction is the most common technique for recognizing moving regions in an image, which takes the gap between the observed image and the reference background.

The following are the key goals of this research:

- Study existing face recognition, identification, and tracking techniques for real-time offline and online applications.
- Develop a method for face recognition and tracking schemes using machine learning methods.
- Face acknowledgment and confirmation in video groupings utilizing a profound learning plan.
- Enhancing an existing CNN-based design for facial recognition and authentication step-by-step.
- The deep fake detection model is designed to identify counterfeit motion detection from Video Surveillance Systems.

2 Literature Survey

Object recognition in video observation frameworks has been an open exploration field for academicians, scholastics, and industry over the course of the past 10 years. Video keep an eye on spear frameworks are security devices that help us in observing different moving articles. Because of the fast development of different kinds of item location, observing, strange way of behaving, and psychological oppressor exercises, the programmed video reconnaissance system is an extremely complete and flourishing area of examination. Video surveillance is commonplace and commonly used for safety and protection through a digital camera monitoring device. Moving object detection (folks, cars, and trucks) and higher-level analysis are the foundations of every smart surveillance device.

Based on the Surveillance camera, Segmentation and Pre-processing, Extraction, and selection of features, and Face detectors, we have surveyed a few research papers and found challenges addressed in this chapter.

S. Jain and Co. [3] proposed a Deep learning hashing method for face detection and its location. L. Zhao et al. [4] developed a reliable system for detecting faces in different lighting and camera angles. T.P. Nguyen et al. [5] proposed a face recognition method that combines a learning framework. M. Hori et al. [6] identified a major challenge of facial recognition as the large appearance variations caused by factors such as viewpoints.

In [7] and [8], variety of feature extraction models developed. They have worked on Face Tracking system. In [9] fostered a remarkable structure for keeping up with recognizable proof that groups and interfaces the essences of various people in expanded video successions. Lin et al. [10] expressed in their paper that they introduced a nondescript framework for checking and following human personalities in an unstructured video. Lin et al. [11] examined that most existing identification calculations perceive the bounding box's area, bringing about foundation clamor in the facial attributes and low recognition exactness. Summary of existing techniques is given in table 1

Table 1. Summary of Existing Techniques

Ref. No	Topics covered	Methodology	Drawbacks
[3]	Face Detection	One-stage approach. Anchor free methods	No motion and fake detection
[4]	Face Detection	YCbCr color model	Monitoring the scene
[5]	Facial Detection	SVM	Local features
[6]	Face tracking	STM, LSTM	Monitoring
[8]	Face tracking	3D model	Online tracking
[9]	Face identification	Hidden Markov model	Efficient coordinates
[10]	Human tracking	Multi face tracking	Real-time detection
[11]	Human tracking	R-CNN, ResNEt-101	Monitoring online

2.1 Research Challenges

Prepossessing and real-time object detection are crucial components of a video surveillance system. It is a turning point for scientists when we create or pick a reasonable and successful technique for a specific application from the immense range of handling and item discovery strategies accessible. In the present case, almost every reconnaissance gadget utilizes a low-goal camera to catch video. Face discovery means isolating the facial locales from the remainder of the picture. It is vigorously dependent on resulting cycles like face acknowledgment and observing.

3 Methodology and Implementation

In this chapter, we have highlighted the methods created for face detection in Realtime processes using IoT Surveillance data using Deep Learning. Below Fig.1 shows the function of the model.

Our research focuses on designing and implementing efficient methods for video surveillance face detection and recognition. Due to the limited image resolution in video surveillance, relatively few modern digital applications and video combinations, such as video investigation, are available. Design changes to the camera's firmware and other settings are often more challenging. The facial recognition calculation is used to detect the presence of a face in a video, and the size of the face determines the accuracy of the location.

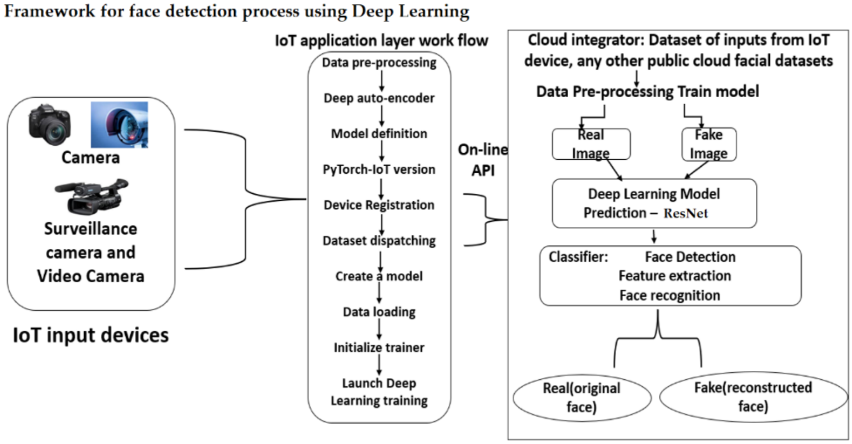
Framework for face detection process using Deep Learning
Camera
Surveillance camera and Video Camera
IoT input devices
IoT application layer work flow
Data pre-processing
Deep auto-encoder
Model definition
PyTorch-IoT version
Device Registration
Dataset dispatching
Create a model
Data loading
Initialize trainer
Launch Deep Learning training
On-line API
Cloud integrator: Dataset of inputs from IoT device, any other public cloud facial datasets
Data Pre-processing Train model
Real Image
Fake Image
Deep Learning Model Prediction – ResNet
Classifier: Face Detection
Feature extraction
Face recognition
Real(original face)
Fake(reconstructed face)

Fig. 1. Deep Learning frame work for face detection

We have introduced a method for improving the efficiency of visual surveillance systems in this paper. Most current face recognition methods are applied to still images. Still, the increasing need for security applications necessitates the development of video-based surveillance systems that can detect, monitor, and recognize multiple faces. Fig 2 shows the process of identification of face motion.

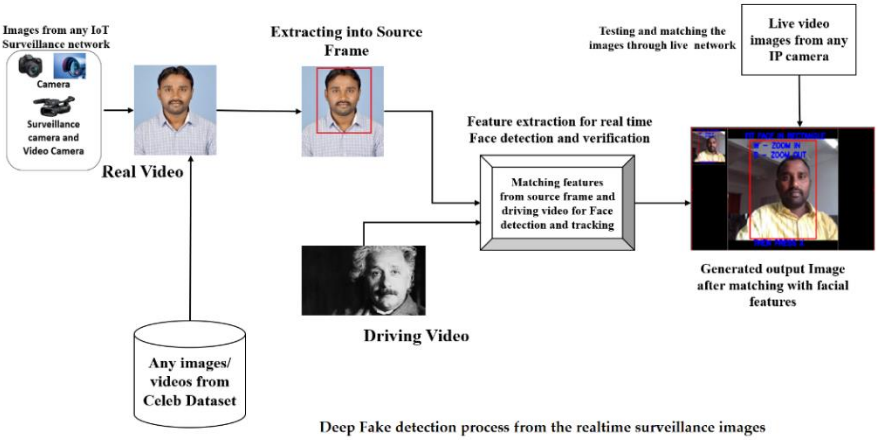
Images from any IoT Surveillance network
Camera
Surveillance camera and Video Camera
Real Video
Extracting into Source Frame
Testing and matching the images through live network
Live video images from any IP camera
Feature extraction for real time Face detection and verification
Matching features from source frame and driving video for Face detection and tracking
Driving Video
Generated output Image after matching with facial features
Any images/ videos from Celeb Dataset
Deep Fake detection process from the realtime surveillance images

Fig. 2. Deep fake detection process from the video surveillance systems

For face detection and tracking, we used a Kalman filtering-based method, followed by building a combined feature extraction model, which was then used to generate a trained database. A Bayesian learning scheme is developed to recognize the detected face. This research can include a CNN-based network model for faster identification in occlusion and low-light situations.

The Bayes filtering has two stages, as discussed in the previous section: prediction and measurement update. A state transition model (also known as a motion model) that describes the state transition of the prior state is needed for prediction xt-1 to the current state xt. We will need a measurement model to update measurements, which defines the mathematical relationship between the current state xt and current height zt. Both the state transfer model and the measurement model for the Linear Kalman Filter (LKF) are linear functions, which can be written as:

Algorithm: Kalman Filter.

$Transition\ model: x_t = Ax_{t-1} + Bu_t + \epsilon_t$ (1)

$Measurement\ model: z_t = Cx_t + \delta_t$ (2)

Initialization: $\mu_0 = E(x_0), P_0 = E\left((x_0 - \mu_0)(x_0 - \mu_0)^T\right)$ (3)

For $t = 1, \dots, \infty$:

Generate sigma points for prediction: $\mathcal{X}_{t-1} = \left[\mu_{t-1}\ \ \mu_{t-1} + \gamma\sqrt{P_{t-1}}\mu_{t-1} - \gamma\sqrt{P_{t-1}}\right]$ (4)

Prediction: $\mathcal{X}_{t|t-1} = g(u_t, \mathcal{X}_{t-1})$

$\hat{\mu}_t = \sum_{i=0}^{2L} w_i^m \mathcal{X}_{i,t|t-1}, \hat{P}_t = \sum_{i=0}^{2L} w_i^c \left(\mathcal{X}_{i,t|t-1} - \hat{\mu}_t\right)\left(\mathcal{X}_{i,t|t-1} - \hat{\mu}_t\right)^T + R_t$ (5)

Generate sigma points for measurement update: $\hat{\mathcal{X}}_t = \left[\hat{\mu}_t, \hat{\mu}_t + \gamma\sqrt{\hat{P}_t}\ \hat{\mu}_t - \gamma\sqrt{\hat{P}_t}\right]$ (6)

Measurement update equations: $\hat{\mathcal{Z}}_t = h\left(\hat{\mathcal{X}}_t\right)$

$\hat{P}_{z_t} = \sum_{i=0}^{2L} w_i^c \left(\hat{\mathcal{Z}}_{i,t} - \hat{z}_t\right)\left(\hat{\mathcal{Z}}_{i,t} - \hat{z}_t\right)^T + Q_t, \hat{P}_{x_t z_t} = \sum_{i=0}^{2L} w_i^c \left(\mathcal{X}_{i,t} - \hat{\mu}_t\right)\left(\hat{\mathcal{Z}}_{i,t} - \hat{z}_t\right)^T$

$K_t = \hat{P}_{x_t z_t} \hat{P}_{z_t}^{-1}, \mu_t = \hat{\mu}_t + K_t(z_t - \hat{z}_t), P_t = \hat{P}_t - K_t \hat{P}_{z_t} K_t^T$ (7)

For Local Binary patterns we used LBP process: $LBP_{N,R}(C) = \sum_{n=0}^{N-1} s(I_n - I_c).2^n$ (8)

Rotation invariance process: $$ROR(x,i) = \begin{cases} \sum_{k=1}^{P-1} 2^{k-1} a_k + \sum_{k=0}^{i-1} 2^{P-i+k} a_k \quad i > 0 \\ x,\ i = 0 \\ ROR(x, P+i), i < 0 \end{cases} \tag{9}$$

$$LBP_{P,R}^{ri} = \min\{ROR(LBP_{P,R}^{ri})|i = 0,1,2,\dots,P-1\}, x = \sum_{k=0}^{P-1} 2^k a_k, a_k \in \{0,1\} \tag{10}$$

The Bayesian learning system for facial identification and classification using the Bayesian model is presented in this section. This procedure follows s=f(x1, x2) where x1 and x1 denote the feature pairs as (x1,x2)€ Rn and s€R represents the likeness. We use entirely connected layers of Multi-Layer Perceptron for the training phase (MLPs). Two photos are used in this process, I1 and I2, which produce x1 and x2 features; this learning process's feature loss can be defined as follows:

$$\mathbb{L}_1(I_1, I_2, \Theta) = \frac{1}{2}\left(N(I_1, I_2\Theta) - J(x_1, x_2)\right)^2 \tag{11}$$

$$\mathbb{L}_2(I_1, I_2, y; \Theta) = (1-y) \times e^{\frac{1}{c}(N(I_1,I_2;\Theta)-t} + y \times e^{-\frac{1}{c}(N(I_1,I_2;\Theta)-t} \tag{12}$$

We have introduced a method for improving the efficiency of visual surveillance systems in this paper. Most current face recognition methods are applied to still images. Still, the increasing need for security applications necessitates the development of video-based surveillance systems that can detect, monitor, and recognize multiple faces. For face detection and tracking, we used a Kalman filtering-based method, followed by building a combined feature extraction model, which was then used to generate a trained database. A Bayesian learning scheme is developed to recognize the detected face. A face detection comparative analysis is shown, demonstrating the performance of the suggested method. This research can include a CNN-based network model for faster identification in occlusion and low-light situations.

Facial recognition in videos is a difficult task due to the changes in pose and lighting, as discussed earlier. However, CNN's recent progress has shown that it can provide a viable solution to this problem. Using CNN, many approaches for face recognition have been developed, but the computational performance needed for precision must be taken to the next level. Therefore, we present a novel video face recognition technique that combines context extraction, a faster RCNN with a direct piece for detection, and CNN for recognition. The proposed cycle (BF-RCNN-VFR) uses a Faster RCNN for video-based facial recognition.

$$r(x_i, x_j) = \log\frac{P(x_i, x_j|H_I)}{P(x_i, x_j|H_E)} = x_i^T G x_i + x_i^T G x_i - 2x_i^T R x_i \tag{13}$$

$$\underset{G,B,b}{\operatorname{argmin}} \sum_{i,j} \max\left[1 - y_{i,j}\left(b - (x_i - x_j)^T G(x_i - x_j) + 2x_i^T B x_j\right), 0\right] \tag{14}$$

In previous chapters, we covered video face detection, tracking, and verification with machine-learning techniques. We shift our attention in this chapter to real-time face detection and validation using a CNN-based scheme.

When attempting to solve complex issues, we often add extra layers; The primary layer of picture acknowledgment could figure out how to identify edges; the subsequent layer could figure out how to perceive surfaces; the third layer could figure out how to see curios, and so forth. The normal Convolutional brain network model, in any case, has an anticipated worth limit, which has been found. In this paper, we utilized the ResNet [12] model for continuous face movement recognition.

4 Results and Discussion

This section uses the IARPA Janus Benchmark A (IJB-A) dataset to test the model. The dataset includes a variety of obstacles, including posture, perspective, and illumination variations. The Celebrity-1000 dataset was primarily concerned with the issue of video-based face recognition. The performance of the YouTube face database, built to recognize faces in videos, is reviewed in this section. CASIAWebFace is the training set to check the efficacy of the end-to-end face detection process. The CASIAWebFace data consists of 10575 people with 494,414 face images, each with a different number of images from tens to hundreds, and data augmentation using horizontal flipping. Fig 3. shows the motion detection and the spoof detection real-time execution process. The result Fig 3. generated from the IJB-A, Celebrity-1000, CASIAWebface, and Youtube dataset. Comparison of Performance and execution of CNN models is given in table 2 and Fig 4,

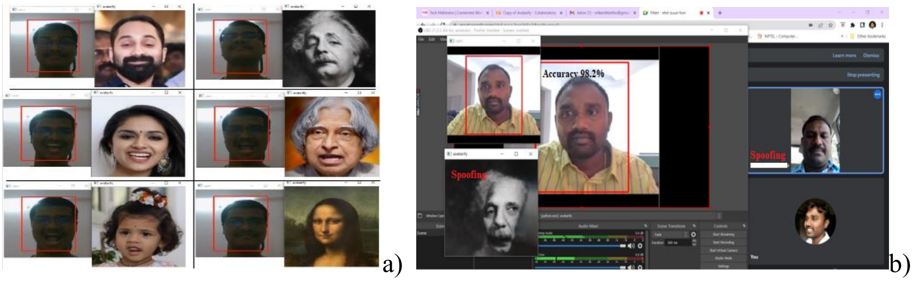

Fig. 3. Result of Face motion and face spoof detection in real-time

Table 2. Performance comparison of CNN models

Model	Dataset used	Performance (%)	Accuracy (%)
AlexNet [13] [14]	CASIA-WebFace, Youtube	89.2	90.1
GoogleNet [13] [14]	IJB-A, Celebrity-1000	88.4	86.8
VGG [13] [14]	Youtube, Celebrity-1000	89.6	89.4
DenseNet [13] [14]	CASIA-WebFace, Youtube	92.5	91.6
ResNet (our model)	Cutomized	96.4	98.6

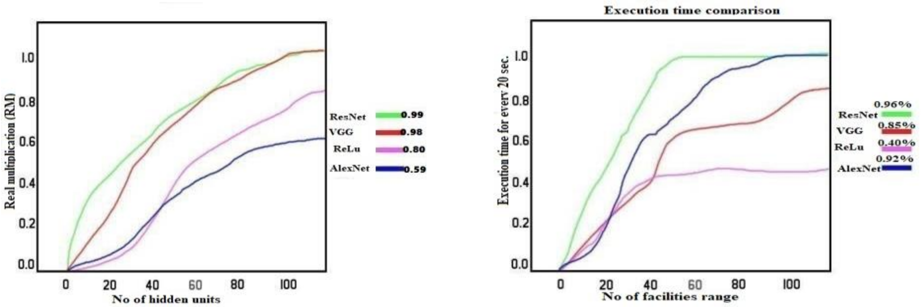

Fig. 4. Performance and Execution comparison of CNN models

5 Conclusion

In this research, we have developed methods for enhancing the efficiency of a visual surveillance system. Most current methods focus on still image-based facial recognition. Still, the growing demand for security applications necessitates the development of video-based surveillance systems that can identify, monitor, and recognize multiple faces. A Bayesian learning scheme is developed to recognize the detected face. Kalman filter is used for feature extraction, and the ResNet model is used for real-time face detection. We have also presented a comparative analysis and demonstrated the robust performance of the suggested method, i.e., the ResNet model, with 98% accuracy. In the Future, we can extend with IoT real-time surveillance systems.

References

1. S. Yousefi, M. T. Manzuri Shalmani, J. Lin and M. Staring, "A Novel Motion Detection Method Using 3D Discrete Wavelet Transform," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 12, pp. 3487-3500, Dec. 2019, doi: 10.1109/TCSVT.2018.2885211.
2. B. Chen, L. Shi and X. Ke, "A Robust Moving Object Detection in Multi-Scenario Big Data for Video Surveillance," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 4, pp. 982-995, April 2019, doi: 10.1109/TCSVT.2018.2828606.
3. S. Jain, A. -M. Creţu, A. Cully and Y. -A. de Montjoye, "Deep perceptual hashing algorithms with hidden dual purpose: when client-side scanning does facial recognition," 2023 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2023, pp. 234-252, doi: 10.1109/SP46215.2023.10179310.
4. L. Zhao, Z. He, W. Cao and D. Zhao, "Real-Time Moving Object Segmentation and Classification from HEVC Compressed Surveillance Video," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 6, pp. 1346-1357, June 2018, doi: 10.1109/TCSVT.2016.2645616.
5. T. P. Nguyen, C. C. Pham, S. V. -U. Ha and J. W. Jeon, "Change Detection by Training a Triplet Network for Motion Feature Extraction," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 2, pp. 433-446, Feb. 2019, doi: 10.1109/TCSVT.2018.2795657.

6. M. A. Abdelwahab, M. Abdel-Nasser and M. Hori, "Reliable and Rapid Traffic Congestion Detection Approach Based on Deep Residual Learning and Motion Trajectories," in IEEE Access, vol. 8, pp. 182180-182192, 2020, doi: 10.1109/ACCESS.2020.3028395.
7. Z. Weng, H. Zhuang, H. Li, B. Ramalingam, R. E. Mohan and Z. Lin, "Online Multi-Face Tracking With Multi-Modality Cascaded Matching," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 6, pp. 2738-2752, June 2023, doi: 10.1109/TCSVT.2022.3224699.
8. C. Siegl, V. Lange, M. Stamminger, F. Bauer and J. Thies, "FaceForge: Markerless Non-Rigid Face Multi-Projection Mapping," in IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 11, pp. 2440-2446, Nov. 2017, doi: 10.1109/TVCG.2017.2734428.
9. Ding, C., &amp; Tao, D. (2015). Robust face recognition via multimodal deep face representation. IEEE Transactions on Multimedia, 17(11), 2049-2058.
10. Lin, C. C., & Hung, Y. (2018). A prior-less method for multi-face tracking in unconstrained videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 538-547).
11. P. J. Lu and J. -H. Chuang, "Fusion of Multi-Intensity Image for Deep Learning-Based Human and Face Detection," in IEEE Access, vol. 10, pp. 8816-8823, 2022, doi: 10.1109/ACCESS.2022.3143536.
12. H. Proença, "The UU-Net: Reversible Face De-Identification for Visual Surveillance Video Footage," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 2, pp. 496-509, Feb. 2022, doi: 10.1109/TCSVT.2021.3066054.
13. Q. He, Z. Mei, H. Zhang and X. Xu, "Automatic Real-Time Detection of Infant Drowning Using YOLOv5 and Faster R-CNN Models Based on Video Surveillance," in Journal of Social Computing, vol. 4, no. 1, pp. 62-73, March 2023, doi: 10.23919/JSC.2023.0006.
14. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1701–1708, 2014.