ResNet-based Gender Recognition on Hand Images

Eren Yildirim

Department of Electronics and Communications Engineering, American University of Malta, Malta | Department of Electrical and Electronics Engineering, Bahcesehir University, Turkiye
eren.yildirim@aum.edu.mt | mustafaeren.yildirim@bau.edu.tr (corresponding author)

Received: 5 September 2024 | Revised: 30 September 2024 | Accepted: 4 October 2024

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.8922

ABSTRACT

The use of biometric features for the surveillance and recognition of certain classes, such as gender, age, and race, is widespread and popular among researchers. Various studies have focused on gender recognition using facial, gait, or audial features. This study aimed to recognize people's gender by analyzing their hand images using a deep learning model. Before training, the images were subjected to several preprocessing stages. In the first stage, the joint points on either side of the hand were detected using the MediaPipe framework. Using the detected points, the orientation of the hands was corrected and rotated so that the fingers pointed upwards. In the last preprocessing stage, the images were smoothened while the edges were preserved by a guided filter. The processed images were used to train and test different versions of the ResNet model. The results were compared with those of some other studies on the same dataset. The proposed method achieved 96.67% recognition accuracy.

Keywords-gender recognition; hand images; guided filter; deep learning; mediapipe; ResNet

I. INTRODUCTION

In the last two decades, the field of automated gender recognition has seen great interest and progress due to its wide range of application areas, including healthcare, surveillance, financial security, and authentication. This progress not only has expanded the methods of gender recognition but has also presented new challenges for further exploration. Early studies on this topic used handcrafted image features, such as Linear Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT), along with conventional Machine Learning (ML) methods, such as Support Vector Machines (SVM), Random Forest (RF), and AdaBoost, to extract information from different parts of the body. Most studies used facial [1, 2], gait [3], fingerprint [4], and palmprint [5] features and body proportions [6].

However, the introduction and rapid advances of Deep Neural Networks (DNNs) changed the research approach. The methods began to employ large dataset collections and training using different deep models. Like conventional methods, researchers collected and focused on different body parts for the training and recognition stages. This study uses a DNN framework for gender recognition from hand images. The primary aim is to recognize the gender of individuals, with a focus on their hand images of both the palm and dorsal sides. There are several studies on this topic, but most of them used hand features such as fingerprints, palmprints, and geometric properties between the knuckles. The main contribution of this study is to improve the hand-based gender recognition task on a biased dataset and analyze the performance of different depth DNNs on the same dataset. Hand features have been used for gender recognition purposes in the last decade. While some of them trained ML models with hand-crafted features, some others used DNNs for training and testing.

In [7], a large dataset called 11k hands was introduced, which consists of images of the dorsal and palmar sides of the hand with ground-truths for gender recognition. This study employed a Convolutional Neural Network (CNN) to extract features to feed an SVM. Another DNN-based study [8] used the Inception model for another dataset, which was relatively small but had variations in posture, achieving sufficient accuracy. In [9], the Global and Part-Aware Network (GPA-Net) was introduced and tested on the 11k hands dataset. This network creates global and local branches on the conv layer to learn robust discriminative global and part-level features. In [10], a multi-CNN model was proposed to detect hand attributes on the 11k hands dataset.

II. PROPOSED METHOD

The proposed method includes several preprocessing steps before training and testing the deep model. These steps include the detection of the joint points on the hand, and orientation correction and smoothing the images while keeping the edges preserved. Figure 1 shows the flow diagram of the proposed method.

A. Detection of Hand Joints by MediaPipe

MediaPipe is one of the many ML frameworks for the detection of hand key points. It can run as web-based, android-based, or as a part of a program written in Python. It works successfully for still images and recorded or live videos [11].

This framework takes a hand image as input and provides 21 key points on the knuckle. Figure 2 shows the key points and their corresponding indexes. Each detected point on the hand has three-dimensional coordinates composed of x, y, and z. The coordinates x and y are normalized between 0 and 1 relative to the width and height of the images, whereas z represents the depth of hand. The depth z decreases as the distance between the hand and the camera decreases. The indexes of the points are independent of the hand orientation, being left-right or the side of the hand facing the camera, such as palm or dorsal. This framework has been used successfully in several studies for different purposes.

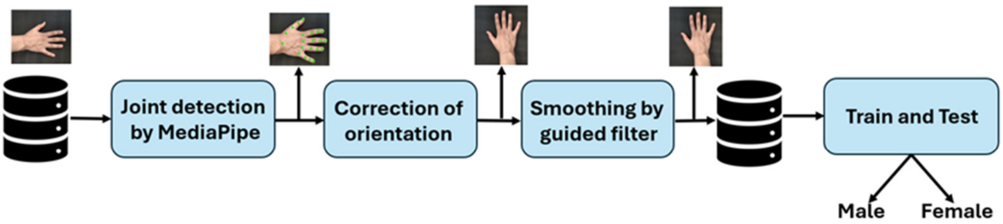
Joint detection by MediaPipe
Correction of orientation
Smoothing by guided filter
Train and Test
Male
Female

Fig. 1. The flowchart of the proposed method.

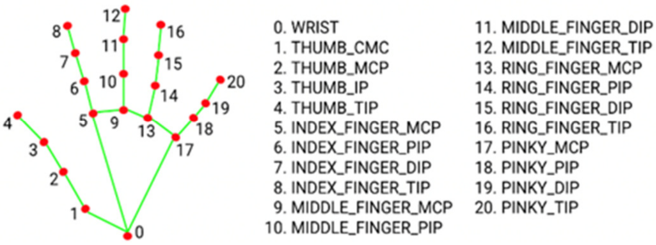

Fig. 2. MediaPipe hand landmark.

B. *Correction of Hand Orientation*

In deep learning, different orientations and rotations of objects of the same class increase the richness of the dataset. This might be advantageous for tests made in uncontrolled mediums, since the dataset has a higher possibility of containing image samples of an object with different rotations. However, this study was conducted for controlled mediums. This means that users are instructed on how to locate their hands in front of the camera. The images have differences in translation, scale, and rotation. The translation and scale in this dataset cannot be utilized. However, the rotation variance can be removed. This can lead to higher accuracy both in the learning and testing processes. Before training, all dataset images were rotated so that the fingers point upward. This was done by calculating the orientation of the hand and rotating it backward around the center. The orientation of the hand is equal to the angle between the points 0 and 9, namely the wrist and the base knuckle of the middle finger. The hand shape is rotated in the opposite direction by the orientation angle obtained.

C. *Smoothing by Guided Filter*

The guided filter [12] is an edge-preserving smoothing technique and an explicit filter. This means that it calculates the output of the filtering process as the weighted average of neighboring pixels of the target image. It takes a guidance image to guide the input image. The guided filter operates under the assumption that a local linear model exists between the guidance image and the filtering result.

The guided filter takes two input images. These are the guidance image I_g and the filtering image I_i. The result of the filtering process I_o is given by

$$I_o(i) = x_k I_g(i) + y_k \quad (1)$$

where y_k and x_k are linear transformation coefficients, i is the pixel location, and k is the window index. The transformation coefficients are calculated as follows:

$$x_k = \frac{\frac{1}{\omega^2}\sum_{i\epsilon\omega_k} I_i(i) I_g(i) - m_k\mu_k}{\sigma_k^2 + \varepsilon} \quad (2)$$

$$y_k = \mu_k - x_k m_k \quad (3)$$

where m_k and σ_k are the mean and the variance of the I_g, and μ_k is the mean of the image I_i in the window ω_k. The level of blurring is controlled by ε in the smoothing process. Since the pixel i is included in multiple windows, the value of the i^{th} pixel in the output image is calculated as:

$$I_o(i) = \bar{x}_i I_g(i) + \bar{y}_i \quad (4)$$

where $\bar{x}_i$ and $\bar{y}_i$ are the expected values of the transformation coefficients among all the windows that include the pixel i. Overall, the guided filter process is controlled by the guidance image, input image, the radius of the window, and ε.

III. EXPERIMENTAL RESULTS

A. *Dataset*

The dataset [13] contains hand images of 100 subjects. The images were taken in a fully controlled environment with a fixed distance from the camera but with different lighting conditions. Moreover, there are large variances in terms of pose, blurriness, and translation. There are 40 images per subject with 10 images for each side of the right and left hands. Five images with errors were excluded, resulting in 3,955 images in total. There are two limitations of the dataset. The first is the imbalance between classes, which might cause a drift in the classification results toward the female class. The second is that the number of subjects is not high enough to generalize the results for people of different races and ages. The images were obtained from different angles with a highresolution device. For personal privacy reasons, only gender and age information was obtained from the people. This dataset is publicly available [13]. Figure 3 shows some images from the dataset. The dataset is unbalanced in terms of gender distribution, as it consists of 1,597 male (40%) and 2,358 female (60%) samples.

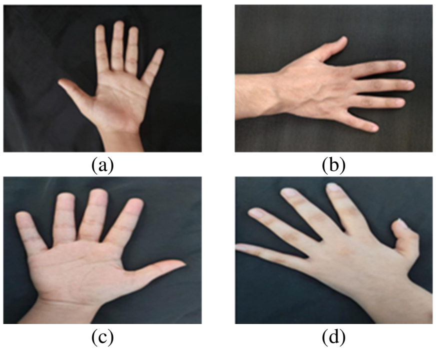

Fig. 3. Sample images of the dataset: (a) male-palm, (b) male-dorsal, (c) female-pam, (d) female-dorsal.

B. *Experiments and Results*

The processed images were provided to the ResNet [14] model for training and testing. The experiment was designed to overcome the limitations of the training stage because it takes a long time and is bounded by the number of layers. The main property of ResNet is its skip connections or its ability to make shortcuts. Moreover, the performance does not decline as the model is getting deeper. Furthermore, its computations are lighter and the ability to train networks is better. The ResNet model skips connections by two to three layers containing ReLU and batch normalization throughout the pipeline [15].

The dataset was divided into 70%, 20%, and 10% for training, testing, and validation, respectively. During the smoothing step, the window radius was taken as 8 and ε was taken as 0.05. The learning rate was determined using the learning rate finder function. This function gives a plot of loss against a range of different learning rates. Also, it outputs several learning rates that give critical points such as the downward slope or the minimum loss. The point with the steepest slope gives the best learning rate. Choosing any other value for the learning rate results in lower accuracy. The learning rates for all models were obtained using the same method. The output of this function is shown in Figure 4. It is a semi-log plot where the x-axis is the learning rate and the y-axis represents the corresponding losses. The recommended learning rate is shown with the valley point on the plot. Therefore, a learning rate of 0.001 was chosen. The dropout rate in the last stage was 50%. 21,252,072 trainable parameters were obtained during training. The reason for using ResNet is that it is a relatively light and simple learning model with sufficient performance. The preprocessing steps were coded in Python. The experiments were also implemented in Python in the Google Colab environment.

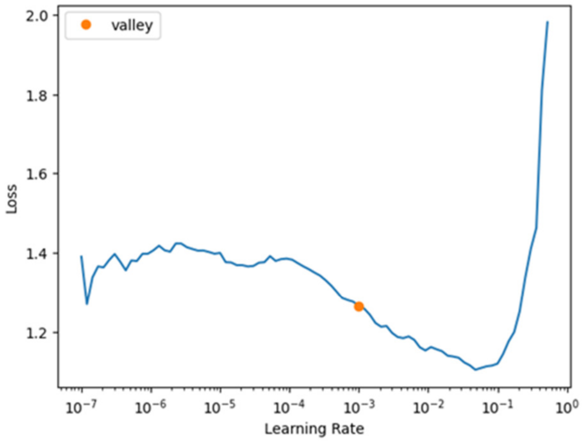

Fig. 4. Loss for different learning rates.

According to multiple tests carried out with the same settings, it was observed that using 10 epochs provided satisfactory results. The highest accuracy was 96.67% for ResNet50. In the tests conducted with more than 10 epochs, the recognition accuracy remained constant or decreased. Therefore, the results obtained with 10 epochs were chosen to compare with the other studies.

TABLE I. RECOGNITION ACCURACY (%)

Method	Accuracy
LBP + Narrow NN [13]	90.00
LBP + Medium NN [13]	91.20
LBP + Wide NN [13]	91.80
LBP + Bilayered NN [13]	90.30
LBP + Tripping NN [13]	88.90
ResNet18	92.74
ResNet34	94.32
ResNet50	**96.67**
ResNet101	94.07

Table I compares the recognition accuracy of the ResNet models in this study along with other studies tested on the same dataset. Comparison was made with both neural network-based studies and all versions of ResNet.

In [13], the results of different algorithmscan be found, namely K-Nearest Neighbor (KNN), SVM, ensemble, decision trees, and naïve Bayesian. According to the results in Table I, ResNet50 outperformed the other models on the same dataset, achieving 96.67% accuracy. The dataset can be considered deep rather than wide. Deeper models are closer to achieving better performance in deep datasets [16]. These results show that the accuracy increases as the depth of the model increases from 18 to 50. However, when the depth of the model increases to 101, a decrease is observed. The most possible reason for this is that the model suffers from overfitting. This overfitting can be overcome using various methods, such as increasing the data or trying different optimizers.

IV. CONCLUSION

This study presented a framework for the task of gender recognition by analyzing hand images. The framework starts with several steps of preprocessing, such as correction of the orientation and smoothing of the images. The dataset consists of hand images, both dorsal and palm, with great variances in translation, size, blurriness, pose, and shape. No further changes, other than rotation and smoothing, were made to the images. The images were examined by training and testing different versions of ResNet, obtaining sufficient results in terms of accuracy.

However, the system has two drawbacks. The first is the degraded accuracy when the pose of the hand is very different from the ones in the dataset. A more generalized and deeper network might overcome this problem. Another solution would be to use a larger dataset with more subjects and variations. The second problem is overfitting when the training model is very deep. This can be fixed by using a larger dataset or trials of different optimizers for regularization.

REFERENCES

[1] F. Bragman, R. Tanno, S. Ourselin, D. Alexander, and J. Cardoso, "Stochastic Filter Groups for Multi-Task CNNs: Learning Specialist and Generalist Convolution Kernels," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 1385–1394, https://doi.org/10.1109/ICCV.2019.00147.

[2] H. Alamri, E. Alshanbari, S. Alotaibi, and M. Alghamdi, "Face Recognition and Gender Detection Using SIFT Feature Extraction, LBPH, and SVM," *Engineering, Technology & Applied Science Research*, vol. 12, no. 2, pp. 8296–8299, Apr. 2022, https://doi.org/10.48084/etasr.4735.

[3] M. E. Yildirim, O. F. Ince, Y. B. Salman, J. K. Song, J. S. Park, and B. W. Yoon, "Gender Recognition using Hog with Maximized Inter-Class Difference," presented at the International Conference on Computer Vision Theory and Applications, Feb. 2016, vol. 4, pp. 106–109, https://doi.org/10.5220/0005715401060109.

[4] G. Jayakala, "Gender classification based on fingerprint analysis," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 10, pp. 1249–1256, 2021.

[5] S. Gornale, A. Patil, and M. Hangarge, "Palmprint Biometric Data Analysis for Gender Classification Using Binarized Statistical Image Feature Set," in *Data Science: Theory, Algorithms, and Applications*, G. K. Verma, B. Soni, S. Bourennane, and A. C. B. Ramos, Eds. Singapore: Springer, 2021, pp. 157–167.

[6] D. T. Nguyen and K. R. Park, "Body-Based Gender Recognition Using Images from Visible and Thermal Cameras," *Sensors*, vol. 16, no. 2, Feb. 2016, Art. no. 156, https://doi.org/10.3390/s16020156.

[7] M. Afifi, "11K Hands: Gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20835–20854, Aug. 2019, https://doi.org/10.1007/s11042-019-7424-8.

[8] R. Mukherjee, A. Bera, D. Bhattacharjee, and M. Nasipuri, "Human Gender Classification Based on Hand Images Using Deep Learning," in *Artificial Intelligence*, Haldia, India, 2022, pp. 314–324, https://doi.org/10.1007/978-3-031-22485-0_29.

[9] N. L. Baisa, B. Williams, H. Rahmani, P. Angelov, and S. Black, "Hand-Based Person Identification using Global and Part-Aware Deep Feature Representation Learning," in *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Salzburg, Austria, Apr. 2022, pp. 1–6, https://doi.org/10.1109/IPTA54936.2022.9784133.

[10] Y. C. Lin, Y. Suzuki, H. Kawai, K. Ito, H. T. Chen, and T. Aoki, "Attribute Estimation Using Multi-CNNs from Hand Images," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Lanzhou, China, Nov. 2019, pp. 241–244, https://doi.org/10.1109/APSIPAASC47483.2019.9023260.

[11] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines." arXiv, 2019, https://doi.org/10.48550/ARXIV.1906.08172.

[12] K. He, J. Sun, and X. Tang, "Guided Image Filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013, https://doi.org/10.1109/TPAMI.2012.213.

[13] E. Aydemir and R. T. E. Alalawı, "Classification Of Hand Images by Person, Age and Gender with The Median Robust Extended Local Binary Model," *Balkan Journal of Electrical and Computer Engineering*, vol. 11, no. 1, pp. 78–87, Jan. 2023, https://doi.org/10.17694/bajece.1171905.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[15] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, "Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer," *Procedia Computer Science*, vol. 179, pp. 423–431, Jan. 2021, https://doi.org/10.1016/j.procs.2021.01.025.

[16] S. H. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," *Neurocomputing*, vol. 378, pp. 112–119, Feb. 2020, https://doi.org/10.1016/j.neucom.2019.10.008.

AUTHORS PROFILE

Mustafa Eren Yıldırım received his BS degree in Electrical Engineering from Bahcesehir University, Istanbul, Turkey, in 2008 and his MS and PhD degrees in Electronics Engineering from the Graduate School of Electrical and Electronics Engineering, Kyungsung University, Pusan, Rep. of Korea, in 2010 and 2014, respectively. He worked as a researcher and lecturer for Kyungsung University until August 2015. Since then he has worked as assistant professor in Turkey and Rep. of Korea until January 2024. Currently he assistant professor in the Department of Electronics and Communications Engineering, American University of Malta. His research interests include machine learning, computer vision, and pattern recognition.