



A Hybrid Feature Fusion Approach for Gender Detection from Palm Image

Bharat M Pipaliya¹, Dr. Jahnvi D Rajyaguru²

¹Bharat M Pipaliya, Department of Computer Science, Atmiya University, Rajkot, Gujarat.

bharat.pipaliya@atmiyauni.ac.in

²Department of Computer Science, Atmiya University, Rajkot, Gujarat. jhanvi.rajyaguru@atmiyauni.ac.in

Abstract

Gender classification from hand images is a useful soft-biometric task in forensic analysis, access control, and demographic analytics, particularly when the face is occluded or unavailable. Existing palm-based methods usually rely either on handcrafted texture descriptors or on a single deep convolutional network, and each family captures only part of the discriminative information in a palm. This paper proposes a hybrid feature fusion framework that combines (i) handcrafted texture and shape descriptors (Histogram of Oriented Gradients, multi-scale Local Binary Patterns, Gabor filter-bank statistics, and geometric palm measurements) with (ii) deep features extracted by a fine-tuned convolutional backbone. The two feature streams are normalized, projected into a shared space, and merged by a gated attention fusion module that learns how much to trust each stream for every input image. A lightweight classifier operates on the fused representation. We describe a subject-independent evaluation protocol on the public 11K Hands dataset, an ablation plan that isolates the contribution of each feature family and fusion strategy, and an explainability analysis based on Grad-CAM. The framework is designed to be reproducible, computationally modest, and robust to illumination and pose variation. Result tables in this manuscript are provided as templates to be completed with the outputs of the accompanying experimental pipeline.

Keywords: gender classification, palmprint, soft biometrics, feature fusion, deep learning, handcrafted features, attention fusion, 11K Hands

Introduction

Soft biometric traits such as gender, age, and ethnicity do not uniquely identify a person, but they narrow down search spaces and support decisions in security, forensic, and human-computer interaction systems [1]. Gender estimation from the face has been studied extensively, yet faces are frequently masked, blurred, or legally restricted as a data source. The human hand, and the palm in particular, offers an alternative cue: it is easy to capture with inexpensive sensors, contains rich ridge, crease, and texture structure, and differs between sexes in both global geometry (palm width, finger-to-palm ratios) and local skin texture.

Although palmprints have been studied thoroughly for identity recognition [2], [3], gender prediction from palm images has received much less attention. Early hand-based studies relied on geometric measurements and classical classifiers [4]. More recent work uses convolutional neural networks (CNNs) and large hand-image collections [5]. Two observations motivate the present work:

1. **Handcrafted and learned features are complementary.** Handcrafted descriptors such as HOG and LBP encode explicit, interpretable local structure (ridge orientation, micro-texture) and are data-efficient. CNN features encode higher-level, learned patterns but can overfit when subject diversity is limited.
2. **Naive concatenation ignores reliability.** Image quality, hand pose, and illumination change from sample to sample. A static concatenation or fixed weighted sum cannot adapt to which feature family is more trustworthy for a given image.

We therefore propose a hybrid feature fusion approach with an input-dependent, gated fusion module.

Contributions

- A hybrid representation combining four handcrafted descriptor groups with deep features from a fine-tuned CNN backbone for palm-based gender detection.
- A gated attention fusion module that learns per-sample weights for the handcrafted and deep streams, compared against early (concatenation), late (score-level), and PCA-reduced fusion baselines.
- A subject-independent evaluation protocol with a full ablation design, cross-hand and cross-side generalization tests, and Grad-CAM-based explainability analysis.
- A discussion of fairness, privacy, and responsible-use considerations specific to gender inference from biometric images.

2. Related Work

2.1 Palmprint analysis

Palmprint recognition has a long history. Zhang et al. introduced online palmprint identification using low-resolution images [2], and Kong et al. surveyed the field, covering principal lines, wrinkles, ridges, and texture-based coding schemes [3]. These works established that palm texture carries strong discriminative information, but their focus was identity rather than demographic attributes.

2.2 Hand-based gender classification

Amayeh et al. used geometry and region-based descriptors of segmented hands for verification and identification [4], showing that hand shape is a stable biometric signal. Afifi released the 11K Hands dataset, comprising roughly eleven thousand dorsal and palmar images from 190 subjects, and reported gender classification and identification results using CNN-based features [5]. Subsequent studies have explored deeper networks and transfer learning, typically treating the task as plain image classification with a single feature source.

2.3 Handcrafted texture descriptors

HOG captures local gradient orientation distributions [6]. Local Binary Patterns encode micro-texture by thresholding neighborhoods and are rotation- and illumination-tolerant in their uniform variants [7]. Gabor filter banks capture orientation- and frequency-selective responses that resemble the receptive fields of early visual cortex and have been widely used for palmprint and fingerprint texture.

2.4 Deep features and transfer learning

Residual networks [8], densely connected networks [9], and compound-scaled networks such as EfficientNet [10] provide strong pretrained backbones. Fine-tuning ImageNet-pretrained backbones is the standard strategy when biometric datasets are relatively small.

2.5 Feature fusion

Fusion in biometrics is commonly categorized by the stage at which information is combined: sensor, feature, score, or decision level [1]. Feature-level fusion preserves the richest information but suffers from the curse of dimensionality and scale mismatch between heterogeneous features. Attention and gating mechanisms offer a learnable alternative to fixed concatenation by weighting streams according to their estimated usefulness for a given input.

2.5.1 Gap addressed

To our knowledge, little prior work on palm-based gender detection combines explicit handcrafted descriptors and deep embeddings under an input-adaptive gating scheme with a controlled ablation study. This paper targets that gap. (Authors should verify this statement against an up-to-date literature search before submission.)

3. Proposed Method

Figure 1 (to be drawn by the authors) shows the pipeline: palm image $\rightarrow$ preprocessing and ROI extraction $\rightarrow$ parallel handcrafted and deep feature streams $\rightarrow$ gated attention fusion $\rightarrow$ classifier $\rightarrow$ gender label.

3.1 Preprocessing and region-of-interest extraction

1. **Hand segmentation.** The palm-side image is converted to the YCbCr color space and skin pixels are segmented by thresholding in the Cb–Cr plane, followed by morphological opening and closing and selection of the largest connected component.
2. **Palm ROI.** Using the finger valley keypoints (between index–middle and ring–little fingers) as reference, a square region covering the central palm is cropped and rotated to a canonical orientation. Where keypoints are unreliable, a hand-landmark detector may substitute.
3. **Photometric normalization.** The ROI is resized to 224×224 , converted to grayscale for the handcrafted stream (RGB retained for the deep stream), and enhanced with Contrast Limited Adaptive Histogram Equalization (CLAHE) to emphasize creases and ridges.
4. **Augmentation (training only).** Small rotations ($\pm 10^\circ$), brightness and contrast jitter, random cropping, and Gaussian blur, applied identically across both streams.

3.2 Handcrafted feature stream

Four descriptor groups are extracted from the preprocessed ROI.

- **HOG.** 9 orientation bins, 16×16 pixel cells, 2×2 block normalization (L2-Hys). Output: vector **h**.
- **Multi-scale uniform LBP.** Radii $R \in \{1, 2, 3\}$ with $P = 8R$ sampling points, computed over a 4×4 spatial grid and concatenated as histograms. Output: vector **l**.
- **Gabor statistics.** A bank of filters at 4 scales and 6 orientations; the mean and standard deviation of each response magnitude, per grid cell, form **g**.
- **Geometric palm measurements.** Palm width, palm length, palm area, aspect ratio, and finger-length-to-palm-width ratios from the segmented hand, forming **m**.

The handcrafted vector is $\mathbf{x}_{\text{hc}} = [\mathbf{h} ; \mathbf{l} ; \mathbf{g} ; \mathbf{m}]$. Each block is z-score normalized using training-set statistics. Optional Principal Component Analysis (PCA) retaining 95% variance reduces dimensionality before projection.

3.3 Deep feature stream

An ImageNet-pretrained backbone (EfficientNet-B0 as the default, with ResNet-50 and DenseNet-121 evaluated in ablations) is fine-tuned on the palm ROI. The classification head is removed, and the global-average-pooled embedding $\mathbf{x}_{\text{dl}}$ (1280-D for EfficientNet-B0) is used as the deep representation. Fine-tuning uses a two-phase schedule: the head and last blocks are trained first with a higher learning rate, then the entire network is unfrozen with a lower learning rate.

3.4 Gated attention fusion

The two streams differ in dimension and scale, so each is first projected into a shared d -dimensional space ($d = 256$):

- $\mathbf{z}_{\text{hc}} = \text{LN}(\mathbf{W}_{\text{hc}} \cdot \mathbf{x}_{\text{hc}} + \mathbf{b}_{\text{hc}})$
- $\mathbf{z}_{\text{dl}} = \text{LN}(\mathbf{W}_{\text{dl}} \cdot \mathbf{x}_{\text{dl}} + \mathbf{b}_{\text{dl}})$

where LN denotes layer normalization. A small gating network observes both projections and produces a scalar weight per stream through a softmax:

- $[\mathbf{a}_{\text{hc}}, \mathbf{a}_{\text{dl}}] = \text{softmax}(\mathbf{W}_{\text{g}} \cdot [\mathbf{z}_{\text{hc}} ; \mathbf{z}_{\text{dl}}] + \mathbf{b}_{\text{g}})$

The fused representation is the gated sum, concatenated with the element-wise product to capture interactions:

$$\bullet \quad \mathbf{z}_f = [\mathbf{a}_{hc} \cdot \mathbf{z}_{hc} + \mathbf{a}_{dl} \cdot \mathbf{z}_{dl}; \mathbf{z}_{hc} \odot \mathbf{z}_{dl}]$$

Because $\mathbf{a}_{hc} + \mathbf{a}_{dl} = 1$, the weights are directly interpretable: they show, per image, whether the model relied more on handcrafted or learned evidence.

3.5 Classification and training objective

The fused vector passes through a two-layer perceptron (256 hidden units, ReLU, dropout 0.4) with a sigmoid output giving $P(\text{female} \mid \text{image})$. Training minimizes binary cross-entropy with label smoothing ($\epsilon = 0.05$), optionally with class weights if the training split is imbalanced:

$$\bullet \quad \mathcal{L} = -[\mathbf{y} \log \mathbf{p} + (1 - \mathbf{y}) \log(1 - \mathbf{p})] + \lambda \cdot \|\mathbf{\theta}\|^2$$

To prevent the gate from collapsing onto one stream, an auxiliary loss is applied to each stream's own linear classifier, so each branch learns to be independently discriminative (deep supervision). The final objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fused}} + \beta (\mathcal{L}_{hc} + \mathcal{L}_{dl})$ with $\beta = 0.3$.

As a further comparison, the fused features can be passed to classical classifiers (SVM with RBF kernel [11], Random Forest [12], XGBoost [13]) instead of the neural head.

3.4.1 Algorithm summary

1. Segment the hand and extract the palm ROI; apply CLAHE.
2. Compute $\mathbf{x}_{hc}$ (HOG, LBP, Gabor, geometry) and normalize.
3. Compute $\mathbf{x}_{dl}$ with the fine-tuned backbone.
4. Project both to the shared space and compute gate weights.
5. Form $\mathbf{z}_f$ and predict the gender probability.

4. Experimental Design

4.1 Dataset

Experiments use the publicly available **11K Hands** dataset [5], which contains images of the dorsal and palmar sides of both hands from 190 subjects aged 18–75, with gender and other metadata provided. Only palmar-side images are used for the main experiments. Dorsal images and cross-side evaluation are reserved for generalization analysis. The dataset's license and usage terms must be followed.

4.2 Subject-independent protocol

A common error in hand-image studies is splitting by image rather than by subject, which leaks identity-specific texture between training and test sets and inflates accuracy. This work splits by **subject**: 70% of subjects for training, 10% for validation, 20% for testing, stratified by gender, so that no person appears in more than one partition. Results are averaged over five random subject-level splits (or five-fold subject-grouped cross-validation), reporting mean $\pm$ standard deviation.

4.3 Implementation details

- Framework: PyTorch for the deep stream; scikit-image and OpenCV for handcrafted descriptors; scikit-learn for classical classifiers.
- Optimizer: AdamW, weight decay $1e-4$, cosine learning-rate schedule, batch size 32, up to 60 epochs with early stopping on validation AUC.
- Learning rates: $1e-3$ (head, phase 1), $1e-4$ (full network, phase 2).
- Seeds fixed and reported; code and configuration files to be released with the paper.

4.4 Evaluation metrics

Accuracy, precision, recall, F1-score, specificity, area under the ROC curve (AUC), and Matthews correlation coefficient (MCC). Because the dataset may be gender-imbalanced, balanced accuracy is also reported. Statistical significance between methods is assessed with McNemar's test and paired bootstrap confidence intervals.

4.5 Baselines and ablations

ID	Configuration	Purpose
B1	HOG + SVM	Classical baseline
B2	LBP + SVM	Texture baseline
B3	Gabor + SVM	Texture baseline
B4	All handcrafted (concatenated) + SVM	Handcrafted-only upper bound
B5	Fine-tuned EfficientNet-B0	Deep-only baseline
B6	Fine-tuned ResNet-50	Backbone comparison
B7	Fine-tuned DenseNet-121	Backbone comparison
F1	Early fusion: concatenation + MLP	Naive feature fusion
F2	Early fusion: concatenation + PCA + SVM	Dimensionality-reduced fusion
F3	Late fusion: average of stream probabilities	Score-level fusion

ID	Configuration	Purpose
P	Proposed: gated attention fusion	Full model
P-a	Proposed without geometric features	Contribution of geometry
P-b	Proposed without auxiliary losses	Contribution of deep supervision
P-c	Proposed without CLAHE	Contribution of preprocessing

4.6 Robustness and generalization tests

- **Photometric robustness:** test-time brightness, contrast, noise, and blur perturbations.
- **Cross-hand:** train on right palms, test on left palms (and vice versa).
- **Cross-side:** train on palmar, test on dorsal images, as a stress test.
- **Data efficiency:** performance with 10%, 25%, 50%, and 100% of the training subjects, to test the hypothesis that handcrafted features help in low-data regimes.
- **Efficiency:** parameter count, model size, and inference time per image on CPU and GPU.

4.7 Explainability

Grad-CAM [14] heat maps are generated for the deep stream to verify that attention falls on the palm region rather than background or imaging artifacts. The learned gate weights (a_{hc} , a_{dl}) are analyzed across image-quality bins to confirm that the fusion module behaves sensibly.

5. Results and Discussion

Note to authors: the tables below are templates. Fill them only with numbers produced by your own runs of the pipeline described in Section 4. Do not report any value you have not measured.

5.1 Main comparison (test set, subject-independent)

Method	Accuracy (%)	F1 (%)	AUC	MCC
B4: Handcrafted + SVM	[]	[]	[]	[]
B5: EfficientNet-B0	[]	[]	[]	[]
F1: Early fusion + MLP	[]	[]	[]	[]
F3: Late fusion	[]	[]	[]	[]
Proposed (gated fusion)	[]	[]	[]	[]

5.2 Ablation study

Variant	Accuracy (%)	Δ vs. full model
Proposed (full)	[]	—
– geometric features	[]	[]
– Gabor features	[]	[]
– auxiliary losses	[]	[]
– CLAHE	[]	[]
Gate replaced by fixed 0.5/0.5	[]	[]

5.3 Expected analysis (to be confirmed by experiments)

The design makes several testable hypotheses, which the results should confirm or refute honestly:

- **H1.** The fused model outperforms both single-stream models, because the handcrafted and deep streams make partly uncorrelated errors.
- **H2.** Gated fusion outperforms static concatenation and fixed-weight averaging, particularly on degraded images, because the gate can down-weight the less reliable stream.
- **H3.** The advantage of handcrafted features grows as training data shrinks.
- **H4.** Accuracy under image-level splitting is markedly higher than under subject-level splitting, demonstrating identity leakage in naive protocols.

If a hypothesis is not supported, that finding should be reported and discussed; negative or mixed results are scientifically valuable.

5.4 Error analysis

Misclassified samples should be reviewed for common causes: low resolution, extreme finger spread or pose, strong shadows, rings or hand accessories, and hand-size outliers. Reporting per-age-group performance is recommended, since age affects skin texture and may change the model's reliance on each stream.

6. Limitations

- **Binary labels.** The task uses a binary female/male label, which does not capture the full range of gender identity. The model predicts a sex-correlated trait from hand morphology, not a person's identity.
- **Single dataset.** The 11K Hands dataset has limited demographic and acquisition diversity. Validation on additional datasets, devices, and populations is needed before any claim of general robustness.
- **Controlled capture.** Images were captured under relatively controlled conditions, so unconstrained, in-the-wild performance may be lower.
- **Age and ethnicity effects.** Hand size and texture vary with age, ethnicity, occupation, and health, which can shift decision boundaries and degrade performance for under-represented groups.

7. Ethical Considerations

Palm images are biometric data. Inferring gender from them can be misused for profiling or surveillance, and any deployment must comply with applicable data-protection law and obtain informed consent. Models should be audited for performance disparities across age and demographic groups, and the system should not be used to make consequential decisions about individuals. This study uses only a public dataset collected with participant consent under its published terms.

8. Conclusion and Future Work

This paper presented a hybrid feature fusion framework for gender detection from palm images that unites interpretable handcrafted descriptors with learned deep embeddings through an input-adaptive gated attention module. The accompanying protocol emphasizes subject-independent evaluation, a thorough ablation design, robustness tests, and explainability analysis. Future work includes extending the approach to multispectral and contactless palm sensing, adding transformer-based backbones, studying self-supervised pretraining on unlabeled palm images, applying domain adaptation across acquisition devices, and combining gender with age estimation in a multi-task setting.

Appendix A: Reference Pipeline (pseudo-code)

for each subject-level split:

```
ROI = extract_palm_roi(image)
ROI = clahe(ROI)
x_hc = concat(hog(ROI), lbp_multiscale(ROI), gabor_stats(ROI), geometry(mask))
x_hc = zscore(x_hc) # train statistics only
x_dl = backbone(ROI_rgb) # fine-tuned CNN embedding
z_hc = LayerNorm(Linear(x_hc))
z_dl = LayerNorm(Linear(x_dl))
a = softmax(Linear(concat(z_hc, z_dl)))
z_f = concat(a[0]*z_hc + a[1]*z_dl, z_hc * z_dl)
p = sigmoid(MLP(z_f))
loss = BCE(p, y) + 0.3 * (BCE(head_hc(z_hc), y) + BCE(head_dl(z_dl), y))
```

References

1. A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, 2004.
2. D. Zhang, W.-K. Kong, J. You, and M. Wong, "Online palmprint identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1041–1050, 2003.
3. A. Kong, D. Zhang, and M. Kamel, "A survey of palmprint recognition," *Pattern Recognition*, vol. 42, no. 7, pp. 1408–1418, 2009.
4. H. Amayeh, G. Bebis, A. Erol, and M. Nicolescu, "Hand-based verification and identification using palm–finger segmentation and fusion," *Computer Vision and Image Understanding*, vol. 113, no. 4, pp. 477–501, 2009.
5. M. Afifi, "11K Hands: gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools and Applications*, vol. 78, pp. 20835–20854, 2019.
6. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, 2005, pp. 886–893.
7. T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.
8. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

9. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE CVPR, 2017, pp. 4700–4708.
10. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.
11. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
12. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
13. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, 2016, pp. 785–794.
14. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618–626.