Enhancing a deep Learning based for human skin segmentation and color tone classification using fast image processing techniques

Zahra Maleki
rosamaleki1382@gmail.com

Ashkan Majidi
Ashkan.Majidi.1382@gmail.com

Milad Hosseini
milad.hosseini532@sharif.edu

Abstract

This paper presents a novel approach to skin segmentation by enhancing MediaPipe's multi-class selfie segmentation model with additional HSV and YCbCr thresholding techniques, and using the output mask for skin color tone classification. We evaluate the effectiveness of these enhancements using metrics such as Mean Intersection over Union (IoU), Mean Precision, Mean Recall, Mean F1 Score, and Mean Dice Score on a diverse dataset. Our results demonstrate significant improvements in skin segmentation accuracy, providing a robust solution for various computer vision applications. Additionally, we examine the impact of Gaussian blur and histogram equalization as preprocessing steps. To further validate our approach, we generated our own dataset and compared the results with those obtained from existing datasets. you can also see the codes in the link https://github.com/rosebaekfany/Skin-Segmentation

1 Introduction

Skin segmentation is a critical task in many computer vision applications, including face recognition, gesture recognition, and human-computer interaction. In this study, we aim to improve the accuracy of skin segmentation by combining the capabilities of MediaPipe's multi-class selfie segmentation model with additional color space thresholding methods in HSV and YCbCr.

2 Related Work

Previous research on skin detection has extensively utilized color spaces like RGB, HSV, and YCbCr to distinguish skin regions from non-skin regions. For instance, (0) developed a skin detection algorithm using correlation rules in RGB, HSV, and YCbCr color spaces. Another study by (0) explored skin detection using these color spaces, demonstrating the effectiveness of color space-based thresholding.

3 Methodology

3.1 MediaPipe

Mediapipe's segmentation model utilizes the Vision Transformer (ViT) architecture with a customized bottleneck and decoder to achieve real-time segmentation of human subjects, particularly focusing on face and body skin. The model processes normalized RGB images at resolutions of [256, 256, 3] or [512, 512, 3], producing a segmentation mask tensor of probabilities across six classes, which is then refined using the Mediapipe image segmenter API. Trained on a dataset of 918 images with 1902 human subjects, the model demonstrated robust performance with a mean IoU of 81.10 (512x512 resolution) and 77.23 (256x256 resolution). While optimized for real-time applications in augmented reality, video conferencing, and entertainment, the model's performance may degrade under challenging conditions such as low lighting or fast motion. It is not intended for life-critical decisions, pixel-perfect tasks, surveillance, or identity recognition. Mediapipe's architecture thus provides an effective and efficient solution for real-time face and body skin segmentation within its defined scope and limitations.

3.2 Color Spaces

Color space is a mathematical model used to represent color information through different components. Various color spaces are employed in applications such as computer graphics, image processing, TV broadcasting, and computer vision. The selection of an appropriate color space is crucial for effective skin detection, influencing the choice

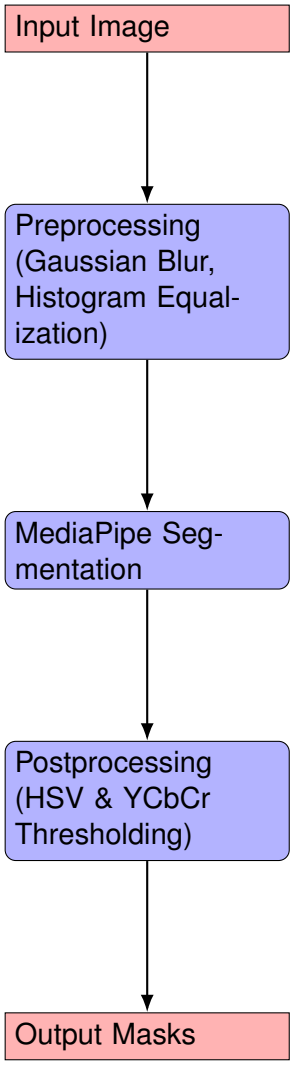

Figure 1: Flowchart illustrating the skin segmentation process.

of thresholds and the performance of the model. In this study, we explore three color spaces: RGB, HSV, and YCbCr.

RGB Color Model: RGB is the default color space for digital images, used extensively in computers, graphics cards, and monitors. It consists of three primary colors: red, green, and blue. Any color can be obtained by mixing these primary colors. RGB values can be normalized for improved skin detection under varying lighting conditions. The RGB color space is advantageous for its simplicity and direct representation of colors.

YCbCr Color Model: YCbCr is widely used in digital video compression standards such as JPEG, MPEG, and is preferred in European television studios. It separates luminance (Y) from chrominance (Cb and Cr), making it effective for image and video compression. The YCbCr model reduces redundant color information and enhances skin detection accuracy by focusing on chrominance components. This color space is particularly useful in dynamic environments with complex backgrounds.

HSV Color Model: HSV is more intuitive for human color perception, representing colors in terms of hue, saturation, and value. This model is beneficial for applications involving interactive color selection and is effective in distinguishing skin tones under different lighting conditions. The hue component ranges from 0° to 360°, covering the entire color spectrum. HSV is particularly useful when precise color differentiation is required, making it suitable for controlled environments.

3.3 Image Preprocessing

Histogram Equalization:

Histogram equalization is employed to enhance the contrast of the images. This technique redistributes the intensity values of the image to cover the entire possible range, thereby improving the visibility of features and aiding in better segmentation. The process involves converting the input image from the BGR color space to the YUV color space. Histogram equalization is then applied to the Y (luminance) channel, after which the image is converted back to the BGR color space. This preprocessing step ensures that the luminance variations are normalized, which is particularly beneficial for images with varying lighting conditions.

Gaussian Blur:

Gaussian blur is applied to reduce image noise and detail, which can help in improving the segmentation process by smoothing the image. This is achieved by convolving the image with a Gaussian kernel, which results in a blurring effect. The size of the kernel determines the extent of the blur. By applying Gaussian blur, high-frequency noise is reduced, and the edges are softened, leading to more robust skin segmentation.

3.4 Evaluation

Each image is preprocessed and then input into the segmentation model. The predicted mask is compared with the ground truth mask to calculate various evaluation metrics, including Intersection over Union (IoU), Precision, Recall, F1 Score, and Dice Score. These metrics provide a comprehensive assessment of the model's performance in detecting skin pixels.

- **Intersection over Union (IoU)**: This metric measures the overlap between the predicted and ground truth masks. It is calculated as the ratio of the intersection area to the union area of the predicted and ground truth masks.

- **Precision**: This metric measures the accuracy of the predicted skin pixels. It is calculated as the ratio of true positive skin pixels to the total predicted skin pixels.

- **Recall**: This metric measures the ability of the model to detect all actual skin pixels. It is calculated as the ratio of true positive skin pixels to the total actual skin pixels.

- **F1 Score**: This is the harmonic mean of Precision and Recall, providing a single metric that balances both aspects.

- **Dice Score**: This metric measures the similarity between the predicted and ground truth masks, similar to the F1 Score, and is calculated as twice the area of overlap divided by the total number of pixels in both masks.

3.5 Proposed Skin Detection Algorithm

The proposed algorithm converts the image into a 2D matrix with each entry representing a pixel's ARGB, HSV, and YCbCr values. The algorithm applies predefined threshold ranges to identify skin pixels. The RGB values are normalized, and the YCbCr and HSV values are calculated using transformation equations. The algorithm compares each pixel's values against standard skin pixel ranges to determine if a pixel is a skin pixel. The skin detection ranges used are as follows:

- **HSV**: $0.0 \leq H \leq 50.0$ and $0.23 \leq S \leq 0.68$

- **RGB**: $R > 95$, $G > 40$, $B > 20$, $R > G$, $R > B$, $|R - G| > 15$, $A > 15$

- **YCbCr**: $Cr > 135$, $Cb > 85$, $Y > 80$, $Cr \leq (1.5862 * Cb) + 20$, $Cr \geq (0.3448 * Cb) + 76.2069$, $Cr \geq (-4.5652 * Cb) + 234.5652$, $Cr \leq (-1.15 * Cb) + 301.75$, $Cr \leq (-2.2857 * Cb) + 432.85$

The integration of these color spaces ensures robust skin detection across various conditions, enhancing the model's overall accuracy.

3.6 Monk Color Classification

Step 1: Monk Color Definition

A predefined set of skin tones, termed "monk" colors, is established as reference points for classification. Each "monk" color is defined by its RGB values, representing a broad spectrum of human skin tones. These colors serve as the standards against which the skin tones extracted from the images are compared.

Step 2: Image Processing and Feature Extraction

Each mask is processed to extract the RGB values of the skin areas. This is achieved by filtering out all non-skin pixels (those that are black or near-black in color) and calculating the average RGB values of the remaining (skin region) pixels. The result is a mean color value that represents the average skin tone for each masked image.

Step 3: Skin Tone Classification

For classification, the method employs a Mean Squared Error (MSE) calculation to determine the similarity between the extracted average skin tone and each of the monk colors. The MSE is computed for each pair of the average skin tone and monk color, and the monk color with the lowest MSE is selected as the matching category. This step quantitatively assesses which predefined skin tone color most closely matches the skin tone derived from the image.

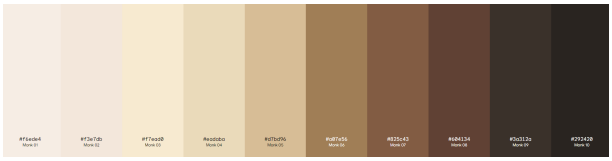

Figure 2: Monk skin scale which contaions 10 number of classes

3.7 Datasets

Pratheepan Dataset:

The dataset used in this study is publicly available for non-commercial use, as detailed by Tan et al., IEEE Transactions on Industrial Informatics (T-II), 2012. It includes images downloaded from Google, captured with various cameras under different conditions. The dataset is organized into four main folders: **FacePhoto**, containing 32 images of single subjects with simple backgrounds; **FamilyPhoto**, comprising 46 images of multiple subjects with complex backgrounds; **GroundT_FacePhoto**, which holds the

ground truth images for the FacePhoto set; and **GroundT_FamilyPhoto**, containing the ground truth images for the FamilyPhoto set. For our evaluation, we exclusively use the **FacePhoto** images, focusing on single faces to assess the Mediapipe segmentation model's performance in controlled conditions. This dataset provides a reliable basis for human skin detection research.

Figure 3: samples of our dataset

Figure 4: YCbCr threshold without MediaPipe on our dataset

Figure 5: HSV Enhancement of our dataset

Sharif Dataset:

To enhance the robustness and comprehensiveness of our evaluation, we gathered the **Sharif Dataset**, consisting of 31 samples that encompass a diverse range of races, ages, and skin tones. This dataset is specifically curated to address the need for evaluating our segmentation methods across various demographic groups, thereby ensuring that the model performs fairly and accurately in real-world scenarios. By incorporating this diverse dataset, we aim to conduct a thorough analysis of skin tone classification and fairness, providing valuable insights into the effectiveness and equity of our segmentation approach.

4 Experiments

We evaluated our enhanced skin segmentation approach on a diverse dataset, using metrics such as

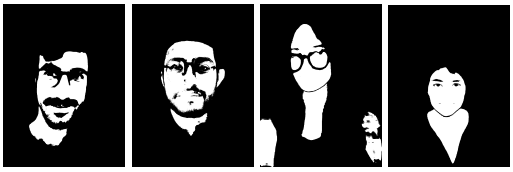

Figure 6: YCbCr Enhancement of our dataset

Mean IoU, Mean Precision, Mean Recall, Mean F1 Score, and Mean Dice Score. Our results are summarized in Table 1. One sample of classification method is in Figure 7

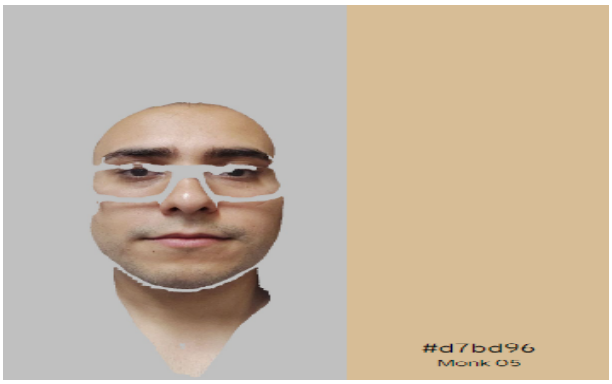

Figure 7: Result of classification method

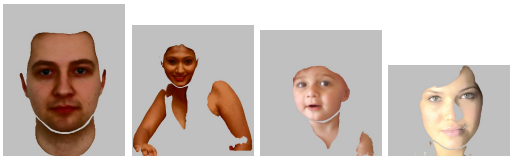

Figure 8: masks using MediaPipe model

5 Conclusion

The combination of MediaPipe's model with additional HSV and YCbCr thresholding enhances skin segmentation accuracy. The analysis of the segmentation methods reveals distinct strengths and weaknesses for each approach. The MediaPipe-only method demonstrates the highest Mean IoU (0.7243) and Mean Recall (0.7703), indicating a robust ability to accurately detect most skin pixels. Its balanced performance across all metrics makes it suitable for general use where no specific condition is particularly challenging. However, it has a lower Mean Precision (0.9280) compared to the enhanced methods, suggesting a higher rate of false positives. On the other hand, the HSV + MediaPipe method boasts the highest Mean Precision (0.9532), ideal for applica-

Method	Mean IoU	Mean Precision	Mean Recall	Mean F1 Score	Mean Dice Score
MediaPipe Only	0.7243	0.9280	0.7703	0.8233	0.8233
MediaPipe + HSV	0.7104	0.9532	0.7392	0.8129	0.8129
MediaPipe + YCbCr	0.6443	0.9727	0.6586	0.7625	0.7625

Table 1: Comparison of Skin Segmentation Methods on Pratheepan dataset

Figure 9: Sample Images for HSV Enhancement

Figure 10: Sample Images YCbCr Enhancement

tions requiring minimal false positives, such as facial recognition in controlled environments. It maintains a competitive Mean IoU (0.7104), ensuring reasonable detection of skin pixels, but it has a slightly lower Mean Recall (0.7392), indicating it might miss some skin pixels, especially in varying lighting conditions. The YCbCr + MediaPipe method excels with superior Mean Precision (0.9727), making it highly accurate in identifying skin pixels and effective in reducing false positives in complex backgrounds or varying lighting conditions, which is advantageous in dynamic environments. However, it suffers from the lowest Mean IoU (0.6443) and Mean Recall (0.6586) among the three methods, indicating it may miss a significant number of skin pixels, and it has lower F1 and Dice scores, suggesting less balanced overall performance. Our approach demonstrates the potential for improved performance in various computer vision applications. Future work will focus on further refining these methods and exploring their application in real-time systems.

Acknowledgments

We thank the developers of MediaPipe and the authors of the referenced papers for their valuable contributions to this field.

References

S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia. 2017. Human Skin Detection Using RGB, HSV and YCbCr Color Models. Thadomal Shahani Engineering College, Sardar Patel Institute of Technology, Mumbai, India.

L. Nanni, A. Loreggia, A. Lumini, and A. Dorizza. 2020. A Standardized Approach for Skin Detection: Analysis of the Literature and Case Studies. University of Padova, University of Brescia, Università di Bologna, Italy.

C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. Google Research.

L. Gustafson, C. Rolland, N. Ravi, Q. Duval, A. Adcock, C.-Y. Fu, M. Hall, and C. Ross. 2022. FACET: Fairness in Computer Vision Evaluation Benchmark. Meta AI Research, FAIR.

W.R. Tan, C.S. Chan, Y. Pratheepan, and J. Condell. 2012. A Fusion Approach for Efficient Human Skin Detection. IEEE Transactions on Industrial Informatics, vol.8(1):138-147 (T-II 2012).