Color Invariant Skin Segmentation

Han Xu[1], Abhijit Sarkar[2], A. Lynn Abbott[1]
[1] Bradley Department of Electrical and Computer Engineering,
[2] Virginia Tech Transportation Institute
Virginia Tech, Blacksburg, VA, USA
{xuhan1996, asarkar1, abbott}@vt.edu

Abstract

This paper addresses the problem of automatically detecting human skin in images without reliance on color information. A primary motivation of the work has been to achieve results that are consistent across the full range of skin tones, even while using a training dataset that is significantly biased toward lighter skin tones. Previous skin-detection methods have used color cues almost exclusively, and we present a new approach that performs well in the absence of such information. A key aspect of the work is dataset repair through augmentation that is applied strategically during training, with the goal of color invariant feature learning to enhance generalization. We have demonstrated the concept using two architectures, and experimental results show improvements in both precision and recall for most Fitzpatrick skin tones in the benchmark ECU dataset. We further tested the system with the RFW dataset to show that the proposed method performs much more consistently across different ethnicities, thereby reducing the chance of bias based on skin color. To demonstrate the effectiveness of our work, extensive experiments were performed on grayscale images as well as images obtained under unconstrained illumination and with artificial filters. Source code: https://github.com/HanXuMartin/Color-Invariant-Skin-Segmentation

1. Introduction

Skin detection refers to the process of identifying pixels that correspond to human skin within image data or video data. Automated skin detection can play an important role for sensitive applications such as face detection and recognition (e.g., [25, 27]), facial expression recognition, gesture recognition [2], content-based image retrieval, filtering of objectionable content [11, 14], skin rendering in computer graphics [4, 10], and virtual reality. Although the last two decades have seen many efforts related to skin detection
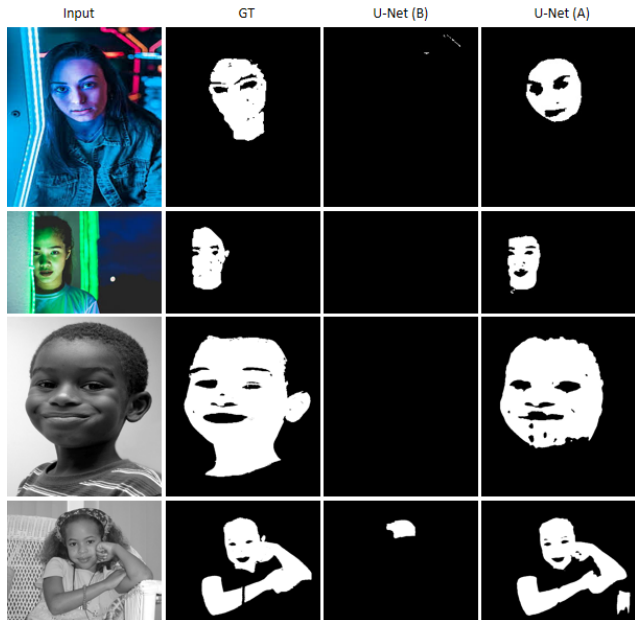

Figure 1. Example results from our skin-detection system, including cases of complex illumination and imaging conditions. *Left to right:* Input image; ground truth; segmentation results using baseline system; results after training using our novel augmentation approach. Dramatic improvements in detected skin regions are apparent for all of these cases.

and skin modeling [20, 46], it is interesting that almost all techniques for image-based skin detection depend heavily on the use of color information. Extensive surveys are provided by Mahmoodi *et al.* [29] and Kakumanu *et al.* [22].

We contend that over-reliance on color cues has imposed performance limits and has also led to bias related to skin tones. A major reason for such bias is dataset imbalance, with the large majority of training samples representing lighter skin tones. Additionally, imaging methods may also introduce variability, including spectral range of sensor arrays (grayscale, near-infrared, RGB) and creative filtering in photography applications (e.g., sepia tones in movies, or

Instagram filters). Examples of such variations are shown in Figure 1.

To address such problems, this paper introduces a new technique for human skin detection that significantly reduces reliance on color information and focuses much more on texture and contextual information to detect the skin pixels in an image. A significant aspect of the system is that color-space augmentation is applied strategically to the training set so that a resulting deep neural network suppresses the system's dependence on color cues. Hence, our high-level strategy has been to guide the training procedure away from color cues and toward features related to visual texture and context. Figure 2 provides an illustration of the augmentation strategy and the training strategy. We demonstrate our procedure by training the U-Net architecture [34] and FCN [28] using ECU [33] dataset and do testing on both ECU and RFW [44] dataset.

The primary contributions of this paper are as follows. *Color invariance.* We describe an approach to automated detection of human skin that does not depend on the color properties of the skin. *Universality.* The resulting system therefore has potential to operate in environments with relatively unconstrained illumination conditions, including extreme cases of over- and underexposed images, grayscale images, and systems that utilize with creative filters (such as Instagram). As such, the system is intended for operation "in the wild," and can relax requirements and reduce costs related to camera selection. *Little or no racial bias.* In our experimental results, we have systematically evaluated the performance of our algorithm for subjects with different skin tones. Using cross-database testing, we have shown that our new algorithm performs virtually uniformly across all of the available annotated skin tones.

Data imbalance is a typical problem in data driven models. Hence, we need to understand the bias before hand and use intelligent algorithm. It is our hope that our color-augmentation strategy for training and testing can be applied widely to other domains, in order to address problems related to racial and social bias.

2. Related work

2.1. Skin detection for natural images

Early approaches to skin detection focused primarily on the use of color cues (e.g., [22, 33, 43]), with the goal of detecting different skin tones under varying illumination conditions. In a few cases, researchers also incorporated cues related to visual texture (e.g., [7, 15, 16]) or shape (e.g., [11]) as a supplement to color information. In one case, researchers considered texture and contextual cues without the use of color [38]. More recently, researchers have applied methods using deep neural networks to the problem of skin segmentation. The different approaches may be grouped loosely into 3 categories: FCN-based [28], R-CNN-based [17, 18], and encoder-decoder models [3, 34].

Under the first category, Zuo *et al.* [47] introduced an end-to-end network for human skin detection by integrating an RNN (Recurrent Neural Network) into an FCN model. They were able to demonstrate improved skin-detection performance in complex environments, including ECU and COMPAQ dataset [21]. He *et al.* [19] later proposed a semisupervised skin-detection method to address the problem of insufficient training samples. Compared to some state-of-art methods. Within the second category, Roy *et al.* [35] used an R-CNN-based approach to reduce the number of false positives by adding a CNN-based skin detector. This approach yielded a substantial improvement over a baseline of using R-CNN only. Nguyen *et al.* [32] integrated a mean-shift hand tracker into Mask R-CNN [18]. They reported improvements of 5% to 9% in detection accuracy, compared to Mask R-CNN alone. Under the encoder-decoder category, Nguyen *et al.* [31] modified the original SegNet [3] architecture by increasing the number of decoders, thereby allowing each encoder to perform multiple tasks at the same time, which discriminate skin components in the hand area more accurately. Topiwala [42] has shown that U-Net stands out among the frequently-used skin detectors on their dataset of the human abdomen with different skin colors, The method based on U-Net was also computationally faster. Tarasiewicz [41] refined the U-Net architecture [34] by considering large-scale contextual features, using inception and dense blocks to reduce occurrences of false positives significantly while doing skin detection.

2.2. Algorithmic bias

This work has been motivated in part by the need to promote demographic fairness in automated systems, particularly relating to differences in skin tones that are related to ethnicity and race. For tasks such as face recognition, techniques have been developed recently to evaluate bias within algorithms and datasets [5], and to improve fairness with respect to such differences [12]. More generally, Mehrabi *et al.* [30] have surveyed the Machine Learning field and have developed a taxonomy of problems that affect bias and fairness within automated systems. Most bias-mitigation systems focus on two types of biases: dataset bias, and task bias. This paper is concerned with the former, which refers to datasets having classes are not represented as well as others within the dataset. Researchers recently have focused on invariant feature learning for protected variables (here, the skin color appearance), and perform database repair to eliminate the representation error [1, 26, 37]. This paper uses the database repair approach through augmentation for de-biasing.
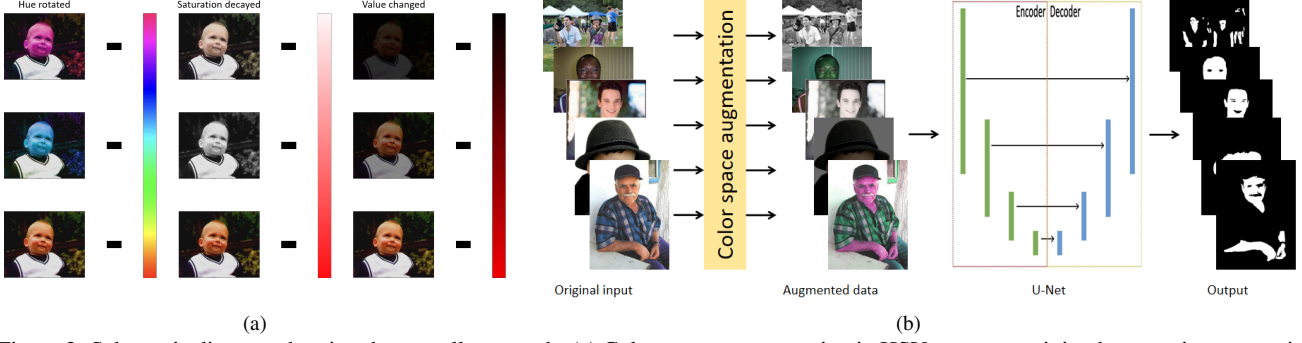
(a)   (b)

Figure 2. Schematic diagram showing the overall approach. (a) Color space augmentation in HSV space containing hue rotation, saturation decay and value change. (b) Overall structure of our skin detection system. During training, each input image undergoes color augmentation.
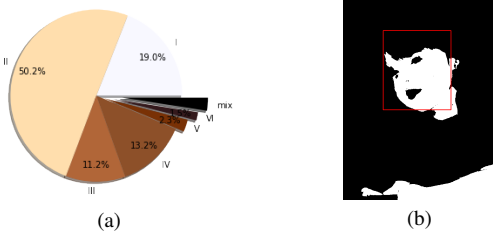

(a)   (b)

Figure 3. (a) Distribution of skin types in the ECU dataset. Labels I-VI refer to the six skin tones described by Fitzpatrick [13]. The group "mix" refers to several skin tone categories in a single image. (b) An example of skin/face evaluation, as described in equation (3).

3. Methods

The overall process is described in Figure 2. We first use an image augmentation approach to create an expanded dataset (Figure 2a), and use the dataset for training a U-Net-based segmentation network for skin detection (Figure 2b).

3.1. Dataset repair by augmentation

In this work, we adopt color-based data augmentations that can add artificial images to mimic alternate representations of the image, in this case, the appearance of the skin. We have implemented augmentation in the HSV (Hue, Saturation, Value) space. In general, the ambient illumination, especially the amount of light reflected by the skin surface, is reflected in the value channel. It includes shadows and exposure. Saturation indicates the purity of the color. On the other hand, the spectral property of the illumination source is reflected in the hue channel and the saturation channel. (Examples are shown in Figure 1.)

Studies have shown that physiological biases, particularly skin tones, can influence computer the development of vision algorithms [5,6,39]. To illustrate a potential cause of bias, Figure 4 provides heatmaps from the ECU testing set. We first classified the those images into six skin-tone categories according to [13]. Then every image in the testing set was converted into HSV color space, and its skin pixels

were allocated into different bins according to the (S, V), (S, H), and (V, H) value pairs. The first six columns show the distributions of skin pixels for the different Fitzpatrick categories in HSV color space. Column 7 in the figure ("W/O") shows the composite distribution for the dataset, and it is clearly seen that the darkest skin tones are poorly represented within HSV space. Our recommended method extends that representation and aims towards a distribution that is much more balanced across all classes. All three rows in column 8 ("W") show improvement.

Figure 5 illustrates how color augmentation works in our experiments. We augmented the training set with three groups of thresholds in H, S, and V channels respectively. Each image in the training set will be converted into fifteen images, so the original training set will be fifteen times larger. This augmented training set will be used for training the skin segmentation models. The thresholds we selected in the experiments are listed in the figure 5.

3.2. Training segmentation networks

The experiment uses NVIDIA GeForce RTX 2070 SUPER GPU with 16 GB GPU memory. The algorithm is trained and tested with U-Net [45] and FCN [9].The U-Net is working under Python 3.8.5 and Tensorflow 2.3.1 environment with no pre-trained models. The FCN is working under Python 3.8.5 and Pytorch 1.70 with a pre-trained VGG-16 network. We train the network with a fixed learning rate of 10^{-4}, and each epoch takes around 79s when batch size is set to 8. The network will use a module ImageDataGenerator [23] in the keras to do data augmentation, which includes image rotation, width shift, height shift, shear, zoom, and horizontal flip with nearest fill mode. We used binary cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(f(y_i)) + (1 - y_i)\log(1 - f(y_i)) \quad (1)$$

where N is the number of segmentation classes. The symbol y_i is the label and $f(y_i)$ is the predicted probability

of the points belonging to the ith class. The original output of the network will be from 0 to 1. Since the pixels should belong to either skin category or non-skin category, we use the function below to make the output $\mathcal{O}$ binary where 1 refers to skin pixels. The threshold δ we set is 0.5. To make the experiments more convincing, we also draw precision-recall curve in the supplementary materials.

$$\mathcal{O} = \begin{cases} 1 & \text{if } f(y_i) \geq \delta \\ 0 & \text{if } f(y_i) < \delta \end{cases} \quad (2)$$

3.3. Datasets

In this work, we have used three datasets. The training and initial evaluation were performed using the benchmark ECU [33] dataset. ECU contains images with diverse attributes including gender, age, skin type, skin-like background, indoor and outdoor images, and images with shadows. The dataset contains 4000 RGB images with manually annotated skin pixels as binary images (see Figure 6 as an example). These images are divided into 1600 images for training, 400 for validation, and 2000 for testing. Note that each of these images is used for color space augmentation (18 total). Hence, we have a total of 30400 images for training. To demonstrate the color invariance of the algorithm, we also transformed the test images to augmented space.

Racial bias is another critical attribute in skin detection systems. To evaluate such bias in our system, we have experimented with six skin types following the Fitzpatrick scale [36]. Figure 3a shows the distribution of images containing individuals with skin types of Type I (less melanin concentration) to Type VI (high concentration of melanin). In the figure, "mix" means that individuals with different skin tones appear in a single image. The figure clearly illustrates the class imbalance within the ECU dataset.

For further evaluation, we selected the RFW (Racial Faces in the Wild) dataset [44] for *cross dataset validation* of our algorithm in order to test whether the proposed algorithm exhibits bias related to skin tone. RFW is a standard test database used to study racial bias in face recognition (see Figure 7 as an example). Four test subsets are provided: Caucasians, Asians, Indians, and Africans. Each subset contains about 3000 individuals and 6000 image pairs for face verification. For our work, the RFW dataset provided 10196 Caucasian faces, 9688 Asian faces, 10308 Indian faces, and 10415 African faces, which shows a good balance across the different groups.

Finally, we created a small dataset of 20 pictures with extreme illumination variations. These images were selected to contain either colored neon illumination or artificial filters, as shown in Figure 1. Then we performed

Table 1. Test results for several skin segmentation methods with the ECU dataset. Our results using U-Net are significantly better than previous methods. For both FCN and U-Net, our use of color-based augmentation improved overall performance of the system.

Methods	Acc	Pre	Rec	F1	IoU
Kolkur *et al.* [24]	83.73	57.00	88.38	69.31	53.03
Dahmani *et al.* [8]	85.95	63.12	77.91	69.74	53.54
Jones *et al.* [21]	89.51	78.23	68.58	73.09	57.59
FCN before aug.	95.78	92.32	86.93	88.66	79.63
FCN after aug.	95.89	92.14	87.70	89.87	81.60
U-Net before aug.	95.59	89.56	**89.15**	89.35	80.76
U-Net after aug.	**96.33**	**92.99**	89.04	**90.97**	**83.44**

manual annotation using SuperAnnotate [40] to identify the skin pixels and non-skin pixels. We conducted extensive testing using this dataset and made pixel-wise evaluations using our ground-truth annotations.

3.4. Evaluation

For the ECU dataset, we used five measures to evaluate the performance: precision, recall, accuracy, F1 score, and IoU. For the RCW dataset, we do not have the ground truth skin annotation. Hence we developed a different method to evaluate the performance. We first used a face detector to get the face area in the image. Then we run the skin detection algorithm trained on the ECU dataset. Next, we identify the skin pixels in the face boundary (as shown in Figure 3b). Finally, we compute the number of skin pixels in the face boundary to the total number of pixels in the face rectangle.

$$Skin/Face = \frac{Skin\ pixels\ detected}{Total\ pixels\ in\ face\ rectangle} \quad (3)$$

As the pose distribution of the image classes in the RCW dataset is uniform, we expect an ideal algorithm to provide a uniform Skin/Face ratio across all ethnicities.

4. Results and Discussion

4.1. Images in the wild

We compared our method with some state-of-the-art skin segmentation systems, including three traditional methods and one FCN based methods. Kolkur *et al.* [24] and Dahmani *et al.* [8] are two thresholding methods which establish some rules in several color spaces to classify a pixel is skin or not. Jones *et al.* [21] is a naive bayes based methods, which predicts the probability of a pixel to be skin after training with given skin masks. The problem behind these traditional methods is the lack of high level features during detection tasks, resulting in the weak robustness against light changes, complex backgrounds or skin color diversity. For both FCN based methods and our U-Net based method, we trained two models, one without
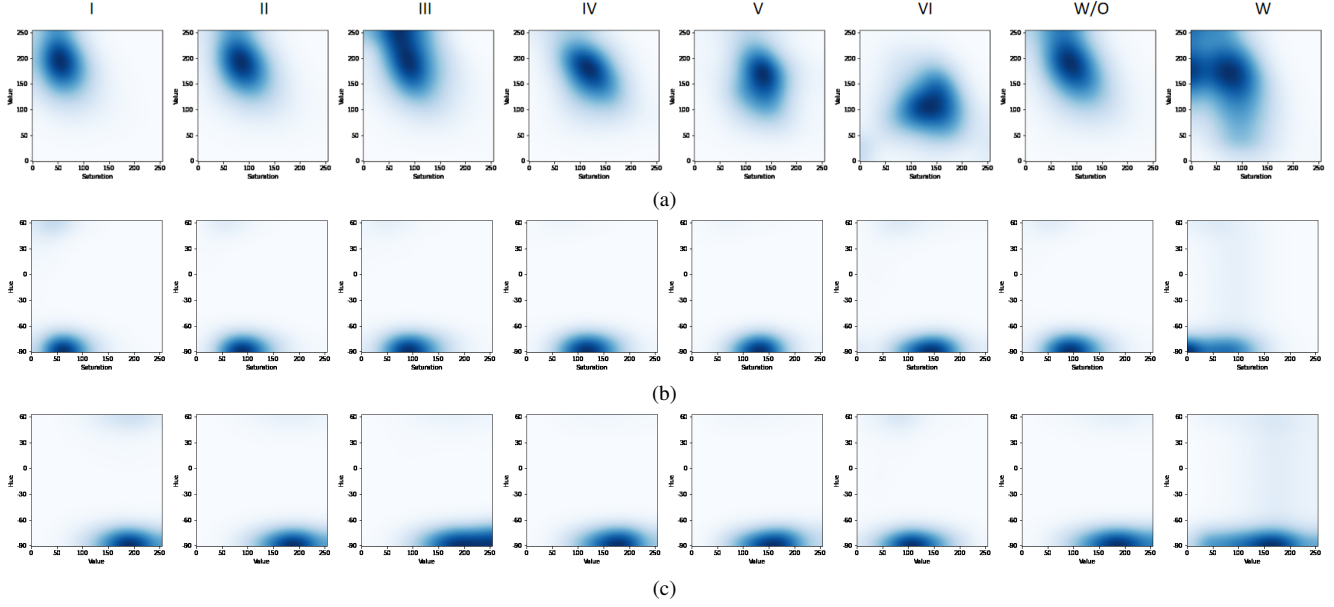
Figure 4. Heatmaps from ECU dataset in three dimensions: (a) Saturation-Value, (b) Saturation-Hue, (c) Value-Hue. The first six columns mark the skin pixel distributions of Fitzpatrick [13] skin tones I-VI. The last two columns refer to the skin pixel distribution of the training set before (W/O) and after (W) our color space augmentation.
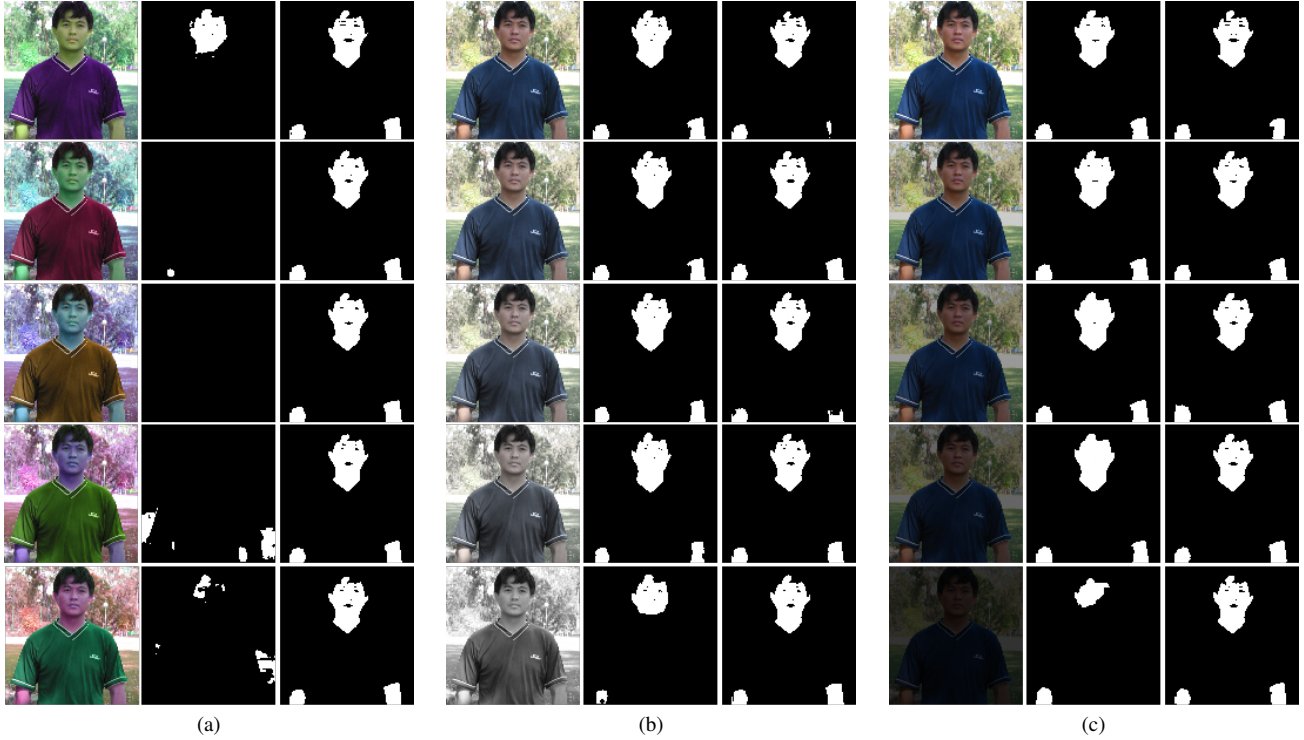

Figure 5. Example of color augmentation across *hue* (a), *saturation* (b), and *value* (c). The first column of each group shows the changed images $\mathcal{I}_{new}$. The second columns show the skin segmentation results without color space augmentation. The third columns show the results with color space augmentation. The input images are rotated at every 60 degrees in the hue channel in the group (a). For group (b), the saturation of images are decayed at ratios of (0.8, 0.6, 0.4, 0.2, 0.0). For group (c), the values of the images are changed at ratios of (1.0, 0.8, 0.6, 0.4, 0.2).

color augmentation and another with color augmentation to confirm the effectiveness of color augmentation.

We first trained the U-Net model with the original RGB images in the ECU dataset (with and without augmentation) and evaluated the performance with the original test set. The precision and recall are shown in Table 1. With
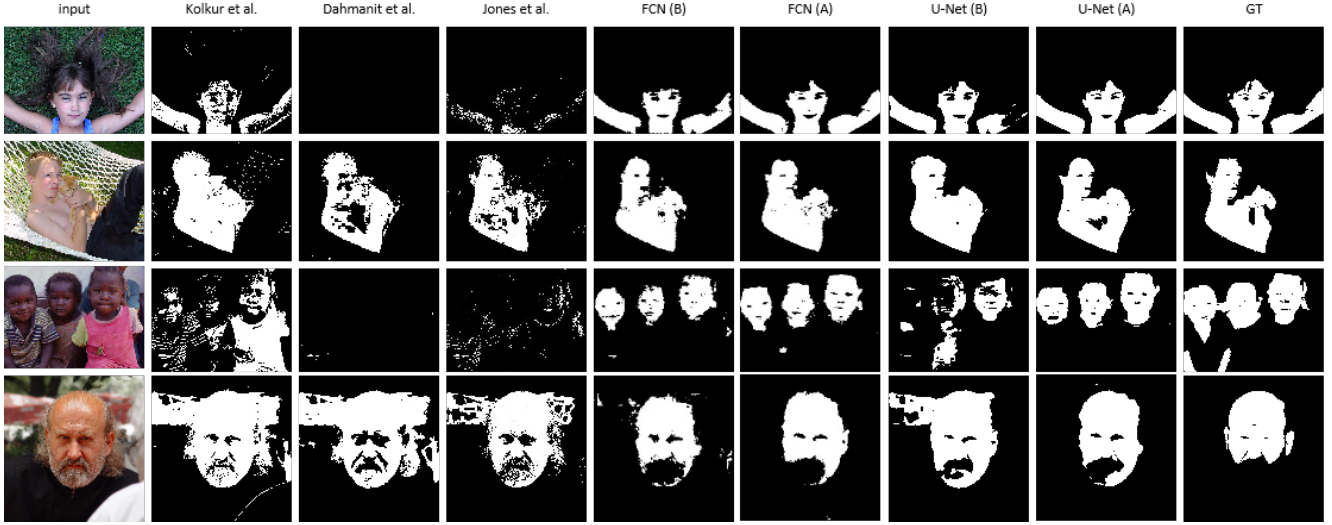
Figure 6. Testing results on the ECU dataset, by various skin segmentation methods including Kolkur *et al.* [24], Dahmani *et al.* [8], Jones *et al.* [21], FCN before (B) and after (A) augmentation, and U-Net before (B) and after (A) augmentation (Columns 2 to 8). Input and ground truth are shown in columns 1 and 9. Our approaches (marked by "(A)") achieve superior results for different backgrounds, genders, poses, and skin tones.

Figure 7. Experimental results from the RFW dataset using several skin segmentation methods. *Left to right:* Kolkur *et al.* [24], Dahmani *et al.* [8], Jones *et al.* [21], FCN before (B) and after (A) augmentation, and U-Net before (B) and after (A) augmentation. Rows 1 to 4 show sample results for the RFW ethnic groups: Caucasian, Asian, Indian, and African.

augmentation, this system yielded a precision of 92.99% and recall of 89.04%, which significantly outperforms the methods of Kolkur *et al.* [24], Dahmani *et al.* [8], and Jones *et al.* [21] (Naive Bayes). The FCN model achieves a precision of 92.14% and a recall of 87.70%. This model also outperforms most CNN-based methods in terms of overall accuracy. While our method shows an accuracy of 96.33%, Tarasiewicz [41] (also a U-Net based architecture) reported an accuracy of 92%.

In Figure 6 we show qualitative comparisons, where the examples cover various skin colors, similar colors in the background, and complex illumination. The first row is a girl lying on the grass with her arms open. The second row is a boy holding a cat, and some of his skin areas are covered by shadow and the cat. The third row is an image of three children with dark skin. The fourth row is a man with a large beard area on his face. These challenging conditions make other methods fail or perform poorly. Dahmani *et al.* [8]

Table 2. F1 scores (%) for the ECU dataset across different skin types. The labels I-VI refer to the six skin tones described by Fitz-patrick [13]. The "mix" column refers to single images containing several individuals with multiple skin categories. The σ column refers to standard derivation of the F1 scores for all columns. Our method with augmentation outperforms in most skin tone categories.

Methods	I	II	III	IV	V	VI	mix	σ
Kolkur *et al*. [24]	67.61	69.96	70.27	70.44	67.61	46.90	72.42	8.14
Dahmani *et al*. [8]	66.10	70.52	71.95	71.01	70.46	56.45	70.45	5.07
Jones *et al*. [21]	64.65	75.89	73.99	74.00	73.28	46.82	77.61	9.99
FCN before aug.	89.03	89.90	90.03	89.56	89.59	83.41	87.37	2.20
FCN after aug.	90.06	90.06	90.34	89.93	90.06	82.98	85.69	2.70
U-Net before aug.	87.16	89.58	90.38	**90.99**	**91.98**	84.72	88.82	2.29
U-Net after aug.	**90.88**	**91.34**	**91.21**	90.55	89.35	**86.05**	**89.60**	**1.84**

Table 3. IoU values (%) for the ECU dataset across different skin types. The column labels are the same as in the previous table. Our method with augmentation outperforms in most skin tone categories.

Methods	I	II	III	IV	V	VI	mix	σ
Kolkur *et al*. [24]	51.07	53.80	54.16	54.36	51.07	30.64	56.76	8.22
Dahmani *et al*. [8]	49.37	54.47	56.19	55.05	54.39	39.33	54.38	5.50
Jones *et al*. [21]	47.77	61.15	58.72	58.72	57.83	30.56	63.41	10.60
FCN before aug.	80.24	81.66	81.87	81.09	81.15	71.54	77.57	3.44
FCN after aug.	81.92	81.91	82.38	81.70	81.91	70.92	74.97	4.22
U-Net before aug.	77.24	81.13	82.44	**83.47**	**85.16**	73.50	79.89	3.68
U-Net after aug.	**83.28**	**84.06**	**83.84**	82.73	80.75	**75.52**	**81.16**	**2.97**

and Jones *et al*. [21] fail in the first and third rows. Kolkur *et al*. [24] classified a large area of background as skin pixels in the fourth row. U-Net (B) works better but still performs poorly in the third row. In contrast, our approach overcame most of the difficulties as stated above, and produced accurate and robust results. (More examples are provided in the supplementary material.)

In order to detect the skin tone bias in the ECU dataset, we further tested the algorithms on different skin tones. Table 2 and Table 3 show that our method outperforms in all categories. Among the three baseline methods, Jones [21] shows the best performance for most of the skin types, but all the methods particularly fail for Type VI (dark skin category). Even deep learning skin segmentation methods show an apparent decline in this dark skin category. Our method consistently shows more than 85% F1 score and more than 75% IoU for all skin types. Moreover, the standard derivations in the last column show that deep learning models have more substantial stability over skin tone bias after color augmentation (more details in the supplementary material).

As shown in Figure 3a, one of the significant problems in the ECU dataset is the imbalance in the images for each skin type. To further understand the robustness of our method, we test with the RFW dataset, which has a balanced dataset across ethnicity. Since RFW dataset does contain manual labels, we only compute the skin/face ratio in this part (see Figure 3 and Function 3). Table 4 shows the evaluated skin/face result of RFW dataset. The model is trained on the ECU dataset using data augmentation. The results show that other methods have different degrees of decline in the "African" group, while our method is stable among

Table 4. Skin/face ratios (%) for the four ethnic groups of the RFW dataset. For the first 3 methods, a significant decline is present for the "African" group. That decline is not present for our FCN and U-Net models, after training with color-space augmentation.

Methods	Cau	Asian	Ind	Afr
Kolkur *et al*. [24]	62.34	62.21	64.31	36.78
Dahmani *et al*. [8]	60.02	59.19	60.95	49.79
Jones *et al*. [21]	47.29	45.49	48.45	20.05
FCN before aug.	67.24	65.78	67.36	64.37
FCN after aug.	73.35	70.59	72.73	73.99
U-Net before aug.	65.47	65.76	69.82	68.39
U-Net after aug.	**77.17**	**73.20**	**76.16**	**78.73**

different races. Our method outperforms the best method for the African group by nearly 29%. Also, our method is significantly better in all the categories compared to the three baselines. Compared the results shown in the last four rows, color augmentation shows its effectiveness on improving the performance of the models. Considering problems of over-prediction, more discussions are shown in the supplementary materials.

Figure 7 shows some qualitative results containing both various skin colors and complex illuminance. The first row is a Caucasian man with background of color similar to his skin. The second row is an Asian woman with one shoulders in the dark. The third row is an Indian man with strong light on his top head. The fourth row is an African woman with her face in shadow. We find that the three baseline methods in columns 2 to 4 are fully confused by the background in the first row. U-Net (B) fails to detect the skin area with intense light in the fourth row. In contrast, our method works better and outputs accurate and complete results.
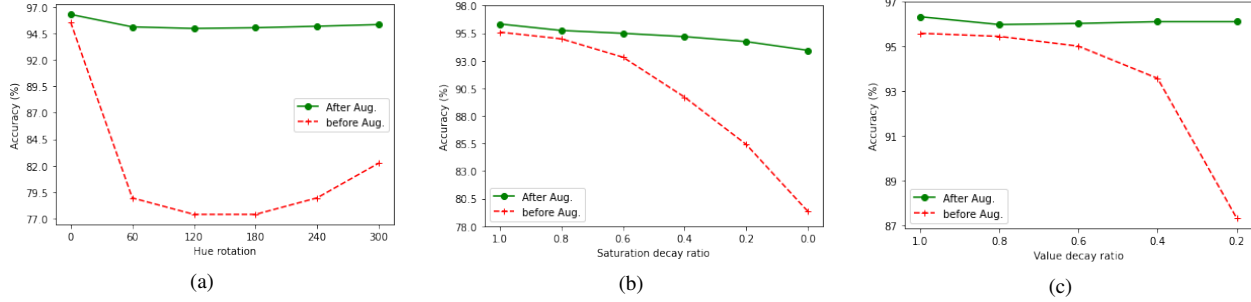
| | (a) | (b) | (c) |

Figure 8. Comparison of U-Net models using color augmentation ("After Aug.") and without color augmentation ("Before Aug.") to test robustness for image filtering in the color space. Augmentation shows their effectiveness for all three dimensions: (a) hue, (b) saturation, and (c) value.

Table 5. Augmentation improves the performance of both U-Net and FCN when tested on images with unconstrained illumination and filters from our self-made dataset.

	IoU before aug.	IoU after aug.	IoU gain by aug.
FCN	12.05	64.92	↑ 52.87
U-Net	11.76	35.85	↑ 24.09

Table 6. Augmentation improves the performance of both U-Net and FCN when tested on grayscale images from the ECU dataset.

	IoU before aug.	IoU after aug.	IoU gain by aug.
FCN	47.13	77.20	↑ 30.07
U-Net	0.55	69.42	↑ 68.87

4.2. Cross dataset testing and ablation study

In order to further test the robustness of the algorithm under different illumination and creative filtering conditions, we ran experiments by testing images transformed by HSV color space augmentation (similar to the training set). We trained two skin segmentation algorithms: One without color augmentation and one with color augmentation. We tested them with the color augmented test set. Figure 8 shows the comparison between the results. For the model without color augmentation, the accuracy falls immediately when tested with images that are modified by hue (Figure 8a), saturation (Figure 8b), or value (Figure 8c). The accuracy remains consistent for the model that was trained with a set of color augmented images. This ablation study shows explicitly the effectiveness of our proposed model.

Figure 5 shows qualitative examples of how the output of our algorithm remains consistent across all the filters, even with drastic changes in the color. Finally, in order to test the performance under ambient light across the spectral, we selected random images from the web and tested them.

Figure 1 shows the robustness of our methods against the drastic illumination changes. The model without color augmentation fails to detect a single skin pixel in the second and third rows, while our method (with augmentation) successfully detects skin pixels in most cases. Qualitative evaluations in Table 5 shows this improvement. IoU increase sharply after color augmentation is applied to the system. More results are in the supplementary material.

4.2.1 Grayscale images

We also conducted tests using grayscale images only. In this part, we used images from the ECU dataset and converted them to grayscale format. We performed testing with the U-Net model and the FCN model, with and without color space augmentation. The results are listed in Table 6.

The table shows that the FCN model without color-space augmentation yields very poor performance, and the resulting IoU is low. For U-Net without augmentation, IoU declines to approximately 0, indicating that the model detected hardly any skin pixels. The performance improves significantly for both of these models, when color augmentation was applied. IoU for the FCN model returned to nearly 80%. Improvements also happen to the IoU for the U-Net model rose to nearly 70% with the help of color augmentation. We list some example results in the Figure 1 and more results can be found in the supplementary material.

5. Conclusion

This paper has introduced a new approach for automated detection of skin in images. Experimental results show that the color-based data augmentation step strategically reduces dependence by the system on color-based cues, and thereby reduces bias related to lightness and color of skin. Using the ECU dataset, our approach has demonstrated better performance than three alternative skin-detection systems. We also conducted an experiment using a racially-balanced dataset, RFW, to illustrate the robustness of our method across different skin tones. Further, the approach addresses problems related to illumination differences (e.g., indoor/outdoor situations, harsh shadows, unnatural lighting) and different sensor parameters (e.g., color, monochromatic, varying spectral sensitivities). We anticipate that similar approaches can be applied more broadly to other Computer Vision tasks.

References

[1] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M. Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2513–2523, 2021. 2

[2] Mohamed Alsheakhali, Ahmed Skaik, Mohammed Aldahdouh, and Mahmoud Alhelou. Hand gesture recognition system. *Information & Communication Systems*, 132–136, 2011. 1

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2

[4] George Borshukov and John P. Lewis. Realistic human face rendering for 'The Matrix Reloaded'. In *ACM Siggraph 2005 Courses*, pages 13–es. 2005. 1

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018. 2, 3

[6] Joy Adowaa Buolamwini. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology, 2017. 3

[7] Oana G. Cula, Kristin J. Dana, Frank P. Murphy, and Babar K. Rao. Skin texture modeling. *International Journal of Computer Vision*, 62(1):97–119, 2005. 2

[8] Djamila Dahmani, Mehdi Cheref, and Slimane Larabi. Zero-sum game theory model for segmenting skin regions. *Image and Vision Computing*, 99:103925, 2020. 4, 6, 7

[9] Yunlong Dong. Trying to be the easiest FCN PyTorch implementation. https://github.com/yunlongdong/FCN-pytorch. 3

[10] Craig Donner, Tim Weyrich, Eugene d'Eon, Ravi Ramamoorthi, and Szymon Rusinkiewicz. A layered, heterogeneous reflectance model for acquiring and rendering human skin. *ACM Transactions on Graphics (TOG)*, 27(5):1–12, 2008. 1

[11] Alexandru F. Drimbarean, Peter M. Corcoran, Mihai Cuic, and Vasile Buzuloiu. Image processing techniques to detect and filter objectionable images based on skin tone and shape recognition. In *Proceedings of the International Conference on Consumer Electronics*, pages 278–279. IEEE, 2001. 1, 2

[12] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020. 2

[13] Thomas B. Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology*, 124(6):869–871, 1988. 3, 5, 7

[14] David A. Forsyth, Margaret Fleck, and Chris Bregler. Finding naked people. *International Journal of Computer Vision*, 1065(1):593–602, 1996. 1

[15] Mehran Fotouhi, Mohammad H. Rohban, and Shohreh Kasaei. Skin detection using contourlet-based texture analysis. In *Proceedings of the Fourth International Conference on Digital Telecommunications*, pages 59–64. IEEE, 2009. 2

[16] Christophe Garcia and George Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1(3):264–277, 1999. 2

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[19] Yi He, Jiayuan Shi, Chuan Wang, Haibin Huang, Jiaming Liu, Guanbin Li, Risheng Liu, and Jue Wang. Semi-supervised skin detection by network with mutual guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2111–2120, 2019. 2

[20] Takanori Igarashi, Ko Nishino, and Shree K. Nayar. The appearance of human skin. Department of Computer Science, Columbia University, 2005. 1

[21] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002. 2, 4, 6, 7

[22] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007. 1, 2

[23] Keras. Image data preprocessing. https://keras.io/api/preprocessing/image/. 3

[24] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia. Human skin detection using RGB, HSV and YCbCr color models. In *Proceedings of the International Conference on Communication and Signal Processing*, pages 324–332. Atlantis Press, 2016/12. 4, 6, 7

[25] J. Kovac, P. Peer, and F. Solina. Human skin color clustering for face detection. In *Proceedings of The IEEE Region 8 EUROCON 2003. Computer as a Tool*, volume 2, pages 144–148, 2003. 1

[26] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019. 2

[27] Qiong Liu and Guang-zheng Peng. A robust skin color based face detection algorithm. In *Proceedings of the 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010)*, volume 2, pages 525–528. IEEE, 2010. 1

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2

[29] Mohammad Reza Mahmoodi and Sayed Masoud Sayedi. A comprehensive survey on human skin detection. *International Journal of Image, Graphics and Signal Processing*, 8(5):1, 2016. 1

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 2

[31] Duong Hai Nguyen, Tai Nhu Do, In-Seop Na, and Soo-Hyung Kim. Hand segmentation and fingertip tracking from depth camera images using deep convolutional neural network and multi-task SegNet. *arXiv preprint arXiv:1901.03465*, 2019. 2

[32] Dinh-Ha Nguyen, Trung-Hieu Le, Thanh-Hai Tran, Hai Vu, Thi-Lan Le, and Huong-Giang Doan. Hand segmentation under different viewpoints by combination of Mask R-CNN with tracking. In *Proceedings of the 5th Asian Conference on Defense Technology (ACDT)*, pages 14–20. IEEE, 2018. 2

[33] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, 2005. 2, 4

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015. 2

[35] Kankana Roy, Aparna Mohanty, and Rajiv R. Sahay. Deep learning based hand detection in cluttered environment using skin segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 640–649, 2017. 2

[36] Silonie Sachdeva. Fitzpatrick skin typing: Applications in dermatology. *Indian Journal of Dermatology, Venereology and Leprology*, 75(1):93, 2009. 4

[37] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019. 2

[38] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. Universal skin detection without color information. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 20–28, 2017. 2

[39] Tomáš Sixta, Julio C. S. Jacques Junior, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. FairFace challenge at ECCV 2020: Analyzing bias in face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 463–481. Springer, 2020. 3

[40] SuperAnnotate. The ultimate training data platform for AI. https://www.superannotate.com/. 4

[41] Tomasz Tarasiewicz, Jakub Nalepa, and Michal Kawulok. Skinny: A lightweight U-Net for skin detection and segmentation. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 2386–2390. IEEE, 2020. 2, 6

[42] A. Topiwala, L. Al-Zogbi, T. Fleiter, and A. Krieger. Adaptation and evaluation of deep learning techniques for skin segmentation on novel abdominal dataset. In *Proceedings of the 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 752–759, 2019. 2

[43] Vladimir Vezhnevets, Vassili Sazonov, and Alla Andreeva. A survey on pixel-based skin color detection techniques. In *Proc. Graphicon*, volume 3, pages 85–92. Citeseer, 2003. 2

[44] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 692–702, 2019. 2, 4

[45] Xuhao Zhi. Implementation of deep learning framework – U-Net, using Keras. https://github.com/zhixuhao/unet. 3

[46] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018. 1

[47] Haiqiang Zuo, Heng Fan, Erik Blasch, and Haibin Ling. Combining convolutional and recurrent neural networks for human skin detection. *IEEE Signal Processing Letters*, 24(3):289–293, 2017. 2