Università degli Studi di Milano Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di laurea in Informatica

Human Skin Detection in Color Images

Relazione della prova finale di:
Michele POZZI
Matricola 845727

Relatore: Prof. Raimondo Schettini
Co-relatore: Dr. Mirko Agarla

Anno Accademico 2020-2021

Scope of Work

The purpose of the thesis is to present a review of the **human skin detection** datasets and approaches of the state of the art, and then perform a comparative in-depth analysis of the most relevant methods on different databases.

Problem Definition

Skin detection is the process of **discriminating skin and non-skin pixels**. It is quite a challenging process because of the large color diversity that objects and human skin can assume and the scene properties (illumination, background, …).

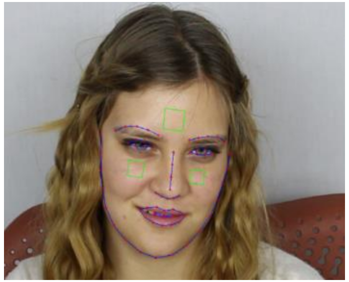
Input image with the subject

Segmented image of the subject’s skin

Problem Definition

Applications:

- **Facial Analysis** [1]
- Gesture Analysis
- **Biomedical** [2]
- Video Surveillance
- Content Filter
- Advertisement



Ramirez et al. 2014 [1]

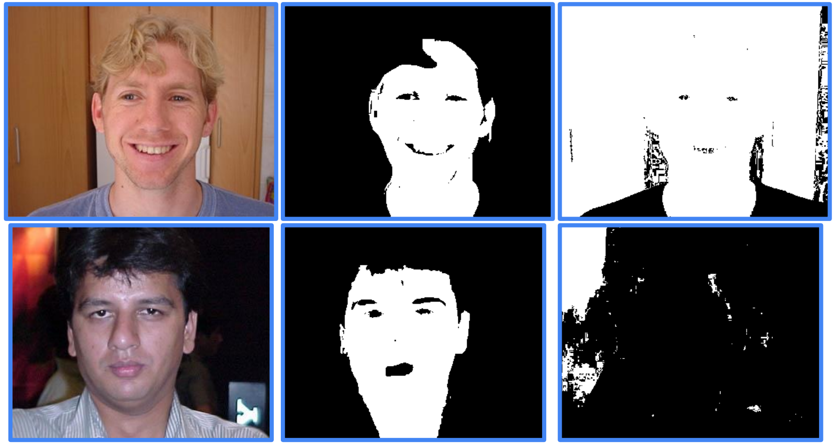
Do et al. 2014 [2]

Limitations:

- Materials with **skin-like colors**
- Wide range of **skin tones**
- **Illumination**
- Cameras color science

Wood color is similar to skin

Lighting changes skin appareance

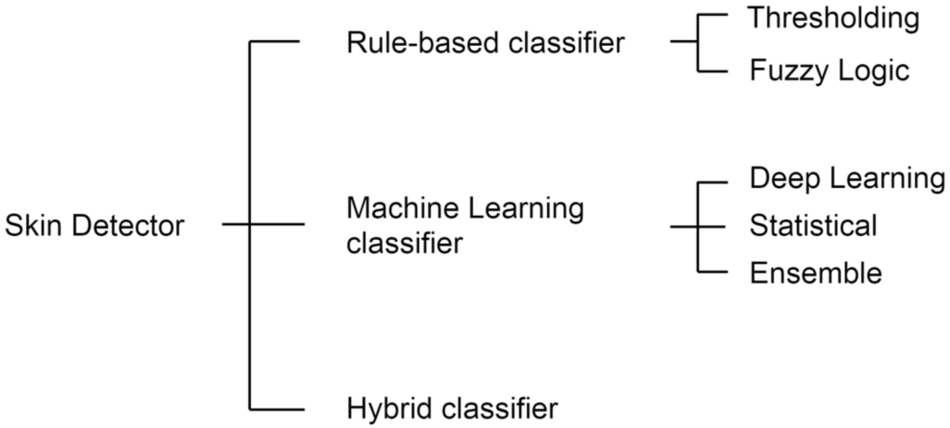


Original image | Ground truth | Prediction

State of the Art

Skin detection is a **binary classification problem**: the pixels of an image must be divided between skin and non-skin classes.

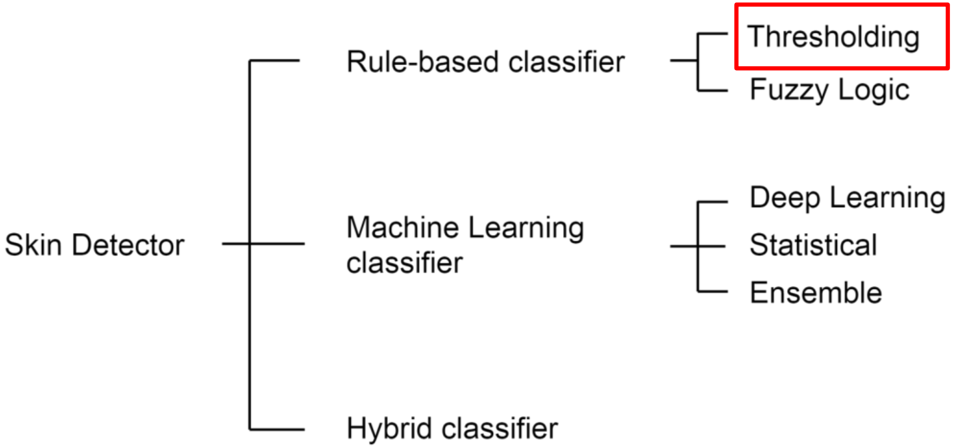
One of several ways to categorize methods is to group them according to how the pixel classification is done.



State of the Art: Thresholding

Skin detection is a **binary classification problem**: the pixels of an image must be divided between skin and non-skin classes.

One of several ways to categorize methods is to group them according to how the pixel classification is done.



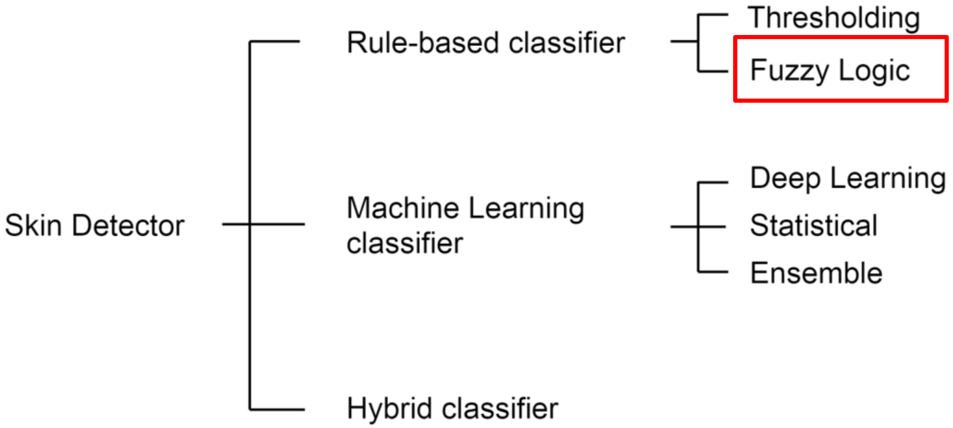
Thresholding approaches use plain rules to classify each pixel as either skin or non-skin.

Example:
(Y,Cb,Cr) is a skin pixel if
133<=Cr<=173
77<=Cb<=127

State of the Art: Fuzzy Logic

Skin detection is a **binary classification problem**: the pixels of an image must be divided between skin and non-skin classes.

One of several ways to categorize methods is to group them according to how the pixel classification is done.

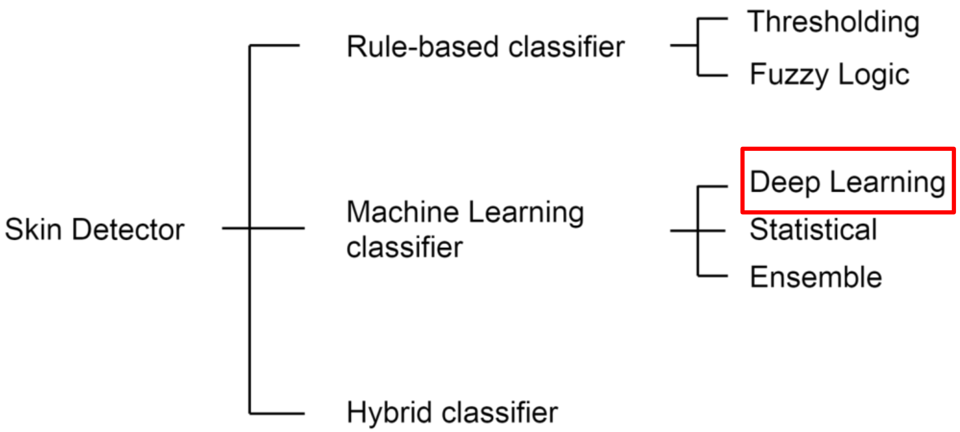


Fuzzy logic approaches use a set of rules to calculate a combined truth value between 0 and 1. The truth value drives the classification.

State of the Art: Deep Learning

Skin detection is a **binary classification problem**: the pixels of an image must be divided between skin and non-skin classes.

One of several ways to categorize methods is to group them according to how the pixel classification is done.

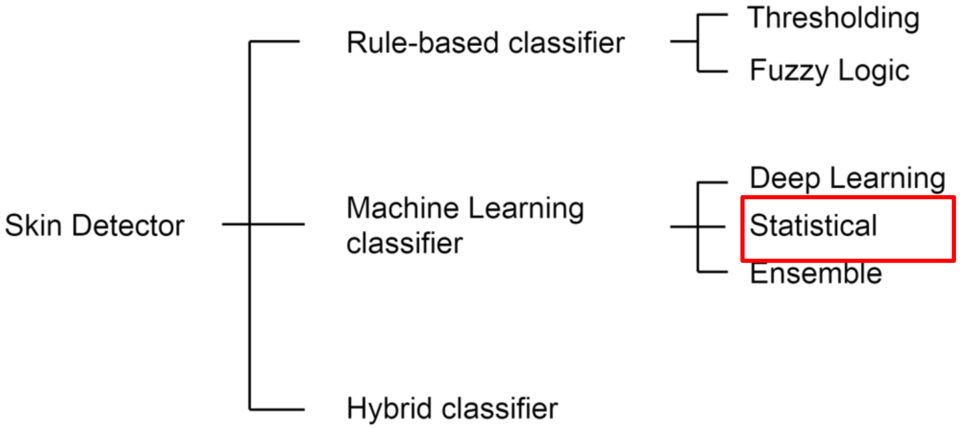


Deep learning approaches use training data to crate a Neural Network model which is then used to perform classification.

State of the Art: Statistical

Skin detection is a **binary classification problem**: the pixels of an image must be divided between skin and non-skin classes.

One of several ways to categorize methods is to group them according to how the pixel classification is done.

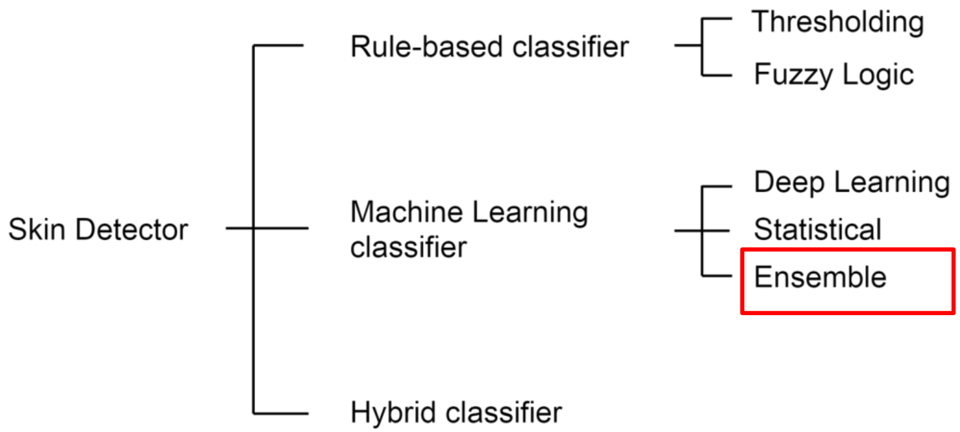


Statistical approaches use training data to crate a statistical model which is then used alongside probability calculus to perform classification.

State of the Art: Ensemble

Skin detection is a **binary classification problem**: the pixels of an image must be divided between skin and non-skin classes.

One of several ways to categorize methods is to group them according to how the pixel classification is done.

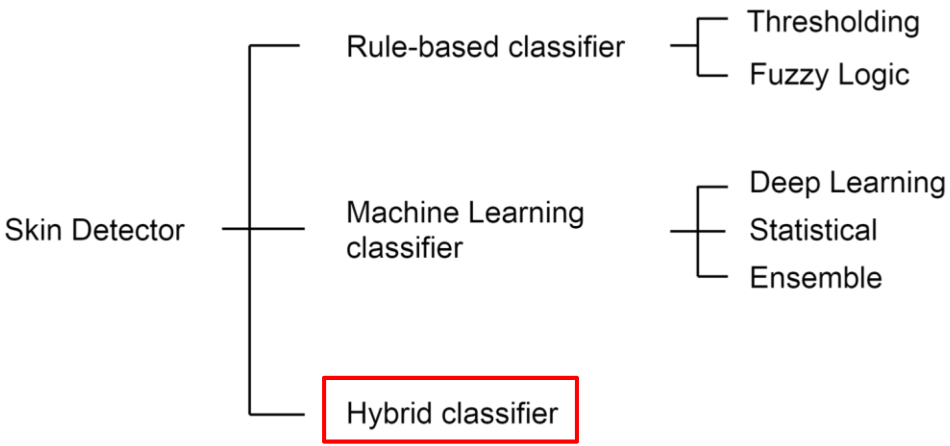


Ensemble approaches use the classifications from different independent machine learning models trained on the same data, as votes for determining the best classification.

State of the Art: Hybrid

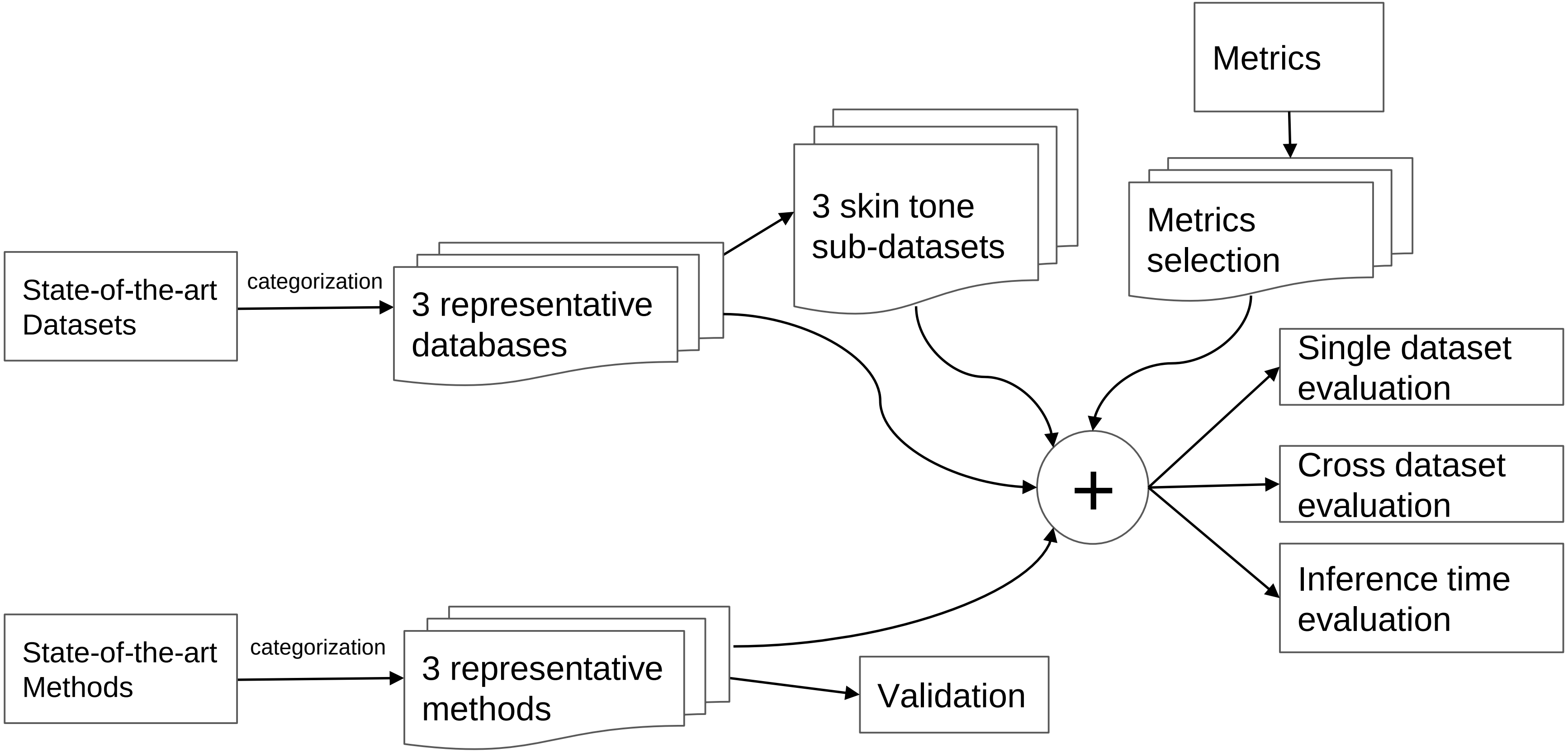
Skin detection is a **binary classification problem**: the pixels of an image must be divided between skin and non-skin classes.

One of several ways to categorize methods is to group them according to how the pixel classification is done.

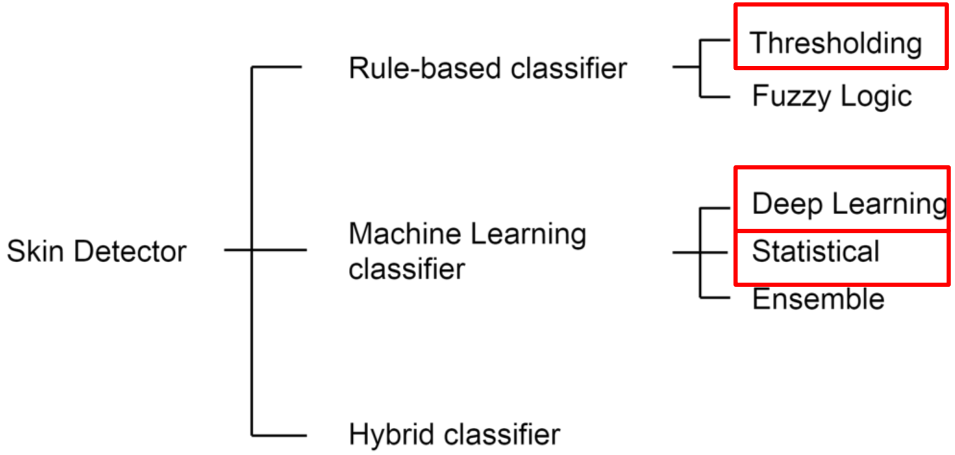


Hybrid approaches make use of different classification techniques <u>that work together</u> to perform the final classification.

Methodological Approach



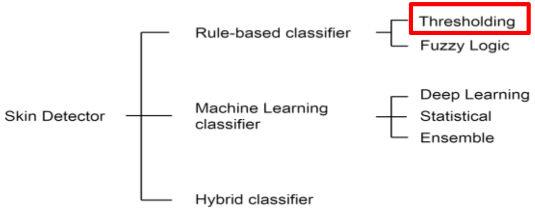
Methodological Approach: Selected methods



A **thresholding** approach has been chosen to demonstrate whether simple rules can achieve powerful results.

A **statistical** and a **deep learning** approaches have been chosen to compare how differently the models behave and generalize, and whether the semantic features extraction capabilities of a CNN can have the upper hand.

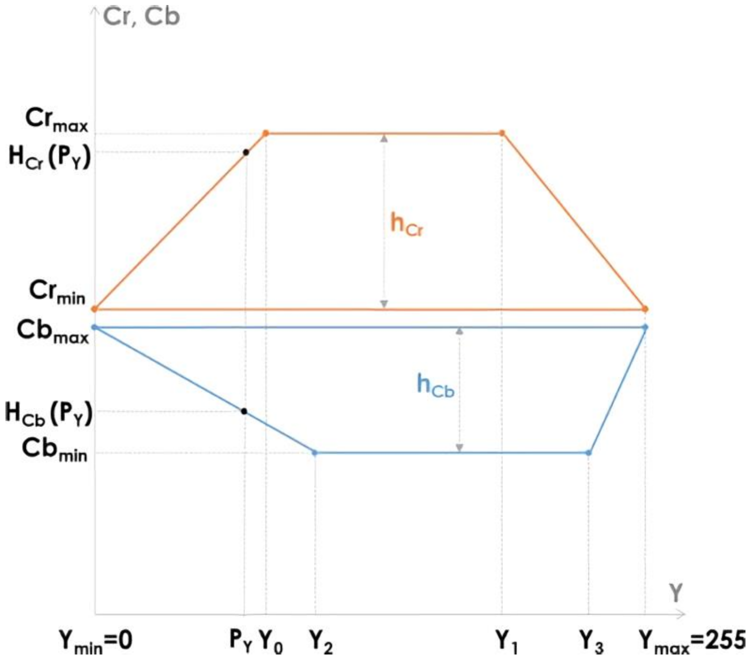
Methodological approach: Rule-based method



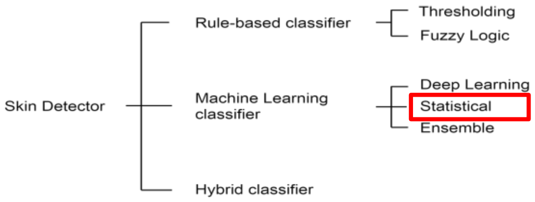
Dynamic Thresholding

Brancati et al. 2017 [3]

1. Input image RGB to YCbCr.
2. Cr_{max} Cb_{min} computation.
3. Pixel-wise computation of the correlation of rules parameters.
4. Pixel-wise correlation rules check.



Methodological approach: Machine learning method



Statistical

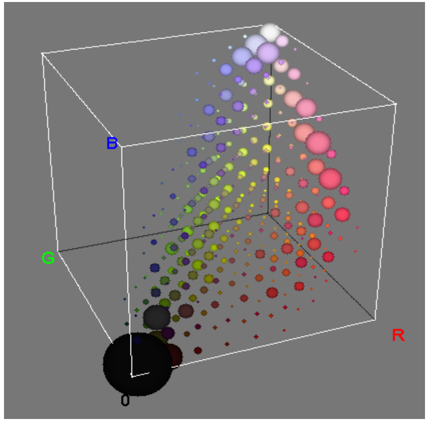
Training:

1. Initialize the skin and non-skin 3D histograms.
2. Pick (*image*, *mask*) from the training set.
3. Loop every *rgb* pixel from *image.*
4. If its mask is a skin pixel, +1 is added to the relative histogram count at coordinates [r,g,b].
5. Return to step 2 until there are images.

Predicting:

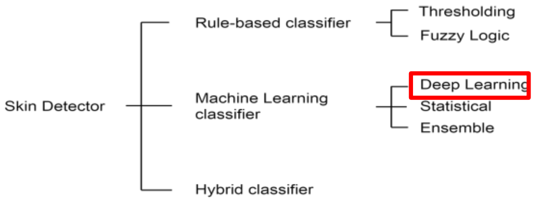
1. Define classifying threshold Θ.
2. Loop every rgb pixel from input image.
3. Calculate rgb probability of being skin.
4. If skin probability > Θ, it is classified as skin.

Original image



3D histogram representation

Methodological approach: Deep learning method

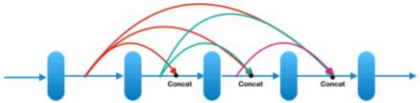


U-Net [4]

Workflow:

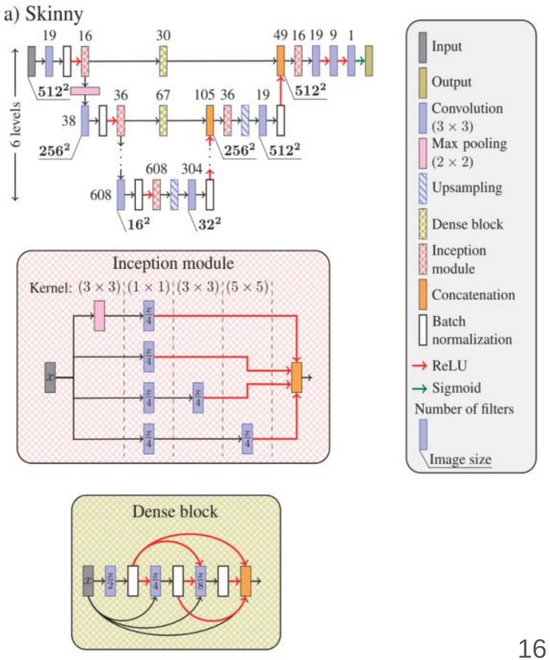
1. Pre-process input image: resize (512×512)px, padding.
2. Extract features in the **contracting pathway** via convolutions and down-sampling, the spatial information is lost while advanced features are learnt.
3. Try to retrieve spatial information through the up-sampling of the **expansive pathway** and the direct concatenations of dense blocks from the contracting pathway.
4. Provide a final classification map.

The salient content size varies between images. **Inception module** combine multiple kernels with different sizes for content adaptation.



Dense block layers are connected in a way that each one receives feature maps from all preceding layers and passes its feature maps to all subsequent layers.

Tarasiewicz et al. 2020 [5]



Datasets

Common datasets used in Skin Detection.

Three datasets have been chosen by considering the popularity, diversity, and size of the databases.

* Skin tones are citations from the papers or eventual labels

Name	Year	No. of Images	Shot Type	Skin Tones
abd-skin	2019	1400	abdomen	african, indian, hispanic, caucasian, asian
HGR	2014	1558	hand	-
SFA	2013	1118	face	asian, caucasian, african
VPU	2013	285	full body	-
Pratheepan	2012	78	full body	-
Schmugge	2007	845	face	skintones labels: light, medium dark
ECU	2005	3998	full body	whitish, brownish, yellowish, and darkish
TDSD	2004	555	full body	different ethnic groups

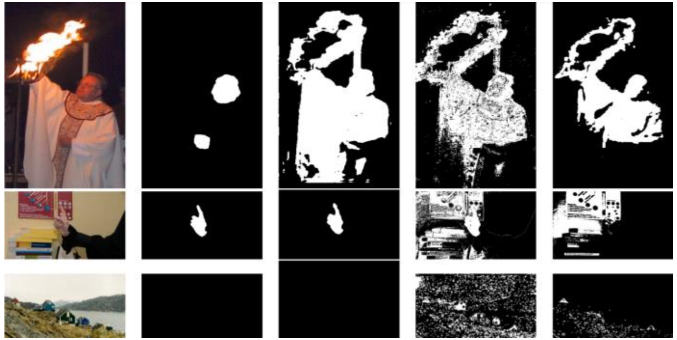
Results: Single dataset evaluation

	Method\Database	ECU	HGR	Schmugge
$F_1 \uparrow$	Deep Learning	**0.9133 ± 0.08**	**0.9848 ± 0.02**	**0.6121 ± 0.45**
	Statistical	0.6980 ± 0.22	0.9000 ± 0.15	0.5098 ± 0.39
	Thresholding	0.6356 ± 0.24	0.7362 ± 0.27	0.4280 ± 0.34
$IoU \uparrow$	Deep Learning	**0.8489 ± 0.12**	**0.9705 ± 0.03**	**0.5850 ± 0.44**
	Statistical	0.5751 ± 0.23	0.8434 ± 0.19	0.4303 ± 0.34
	Thresholding	0.5088 ± 0.25	0.6467 ± 0.30	0.3323 ± 0.28
$D_{prs} \downarrow$	Deep Learning	**0.1333 ± 0.12**	**0.0251 ± 0.03**	**0.5520 ± 0.64**
	Statistical	0.4226 ± 0.27	0.1524 ± 0.19	0.7120 ± 0.54
	Thresholding	0.5340 ± 0.32	0.3936 ± 0.36	0.8148 ± 0.48

Schmugge presents high standard deviations that can be attribuited to its diverse content, featuring different subjects, backgrounds, and lighting.

$$D_{prs} = \sqrt{(1 - PR)^2 + (1 - RE)^2 + (1 - SP)^2}$$

PR - Precision
RE - Recall
SP - Specificity
(1,1,1) - Ideal ground truth



Original Ground truth CNN Statistical Threshold

All approaches struggle on this image. The lighting could be the cause.

Color-based method struggle on images without skin pixels and with materials with similar color.

Results: Cross dataset evaluation

	Training	ECU		HGR		SCHMUGGE	
	Testing	HGR	SCHMUGGE	ECU	SCHMUGGE	ECU	HGR
$F_1 \uparrow$	Deep Learning	**0.9308 ± 0.11**	**0.4625 ± 0.41**	**0.7252 ± 0.20**	0.2918 ± 0.31	**0.6133 ± 0.21**	**0.8106 ± 0.19**
	Statistical	0.5577 ± 0.29	0.3319 ± 0.28	0.4279 ± 0.19	**0.4000 ± 0.32**	0.4638 ± 0.23	0.5060 ± 0.25
$IoU \uparrow$	Deep Learning	**0.8851 ± 0.15**	**0.3986 ± 0.37**	**0.6038 ± 0.22**	0.2168 ± 0.25	**0.4754 ± 0.22**	**0.7191 ± 0.23**
	Statistical	0.4393 ± 0.27	0.2346 ± 0.21	0.2929 ± 0.17	**0.2981 ± 0.24**	0.3318 ± 0.20	0.3752 ± 0.22
$D_{prs} \downarrow$	Deep Learning	**0.1098 ± 0.15**	**0.7570 ± 0.56**	**0.3913 ± 0.26**	**0.9695 ± 0.44**	**0.5537 ± 0.27**	**0.2846 ± 0.27**
	Statistical	0.5701 ± 0.29	1.0477 ± 0.35	0.8830 ± 0.23	1.0219 ± 0.42	0.7542 ± 0.30	0.6523 ± 0.27
$F_1 - IoU \downarrow$	Deep Learning	**0.0457**	**0.0639**	**0.1214**	**0.0750**	0.1379	**0.0915**
	Statistical	0.1184	0.0973	0.1350	0.1019	**0.1320**	0.1308


Original


Ground truth


CNN


Statistical

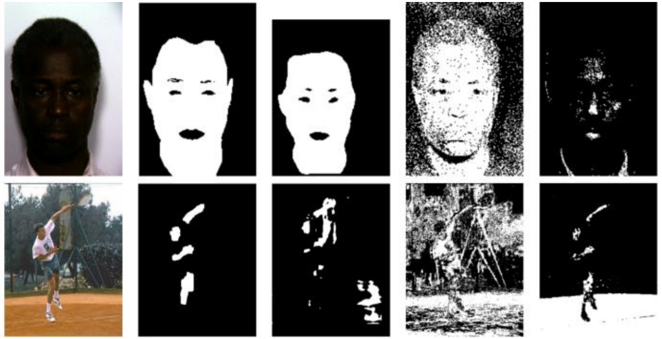
Train: HGR
Test: Schmugge

The statistical approach outperforms the CNN in these cases, but they are both far away from the ideal ground truths.
The statistical approach may have more False Positives as it loses on Dprs and the difference between F1 and IoU.
In this case, the CNN seems to report more False Positives and False Negatives.

The metric **F1 – IoU** is taken into consideration to get a better idea of the number of True Positives compared to False Positives and False Negatives.

Results: Single skin tones evaluation

	Method\Database	DARK	MEDIUM	LIGHT
$F_1 \uparrow$	Deep Learning	$\mathbf{0.9529 \pm 0.00}$	$\mathbf{0.9260 \pm 0.15}$	$\mathbf{0.9387 \pm 0.12}$
	Statistical	0.8123 ± 0.02	0.7634 ± 0.19	0.8001 ± 0.15
	Thresholding	0.2620 ± 0.14	0.6316 ± 0.20	0.6705 ± 0.14
$IoU \uparrow$	Deep Learning	$\mathbf{0.9100 \pm 0.01}$	$\mathbf{0.8883 \pm 0.18}$	$\mathbf{0.9006 \pm 0.14}$
	Statistical	0.6844 ± 0.03	0.6432 ± 0.17	0.6870 ± 0.16
	Thresholding	0.1587 ± 0.10	0.4889 ± 0.19	0.5190 ± 0.14
$D_{prs} \downarrow$	Deep Learning	$\mathbf{0.0720 \pm 0.01}$	$\mathbf{0.1078 \pm 0.21}$	$\mathbf{0.0926 \pm 0.15}$
	Statistical	0.3406 ± 0.05	0.3452 ± 0.23	0.3054 ± 0.20
	Thresholding	0.8548 ± 0.12	0.5155 ± 0.24	0.4787 ± 0.17



Dark skin tone

Medium skin tone

Original Ground truth CNN Statistical Thresholding

The skin tones sub-datasets are taken from the **Schmugge** dataset, which includes labels.

Dark skin tones presented too few images and has been data-augmented to get at least 100 images that look natural.
Trasformations applied:
- Horizontal Flip
- Rotate between -15 and +15 degrees
- Random Crop of 0.8 image size

Dark presents an almost null standard deviation, indicating that the diversity of the images might not be very high.

The thresholding approach struggles to classify dark skin tones.

Results: Cross skin tones evaluation

	Training	DARK		MEDIUM		LIGHT	
	Testing	MEDIUM	LIGHT	DARK	LIGHT	DARK	MEDIUM
$F_1 \uparrow$	Deep Learning	0.7300 ± 0.25	0.7262 ± 0.26	$\mathbf{0.8447 \pm 0.13}$	$\mathbf{0.8904 \pm 0.14}$	$\mathbf{0.7660 \pm 0.17}$	$\mathbf{0.9229 \pm 0.11}$
	Statistical	$\mathbf{0.7928 \pm 0.11}$	$\mathbf{0.7577 \pm 0.12}$	0.5628 ± 0.14	0.7032 ± 0.14	0.5293 ± 0.20	0.7853 ± 0.11
$IoU \uparrow$	Deep Learning	0.6279 ± 0.27	$\mathbf{0.6276 \pm 0.28}$	$\mathbf{0.7486 \pm 0.15}$	$\mathbf{0.8214 \pm 0.16}$	$\mathbf{0.6496 \pm 0.21}$	$\mathbf{0.8705 \pm 0.13}$
	Statistical	$\mathbf{0.6668 \pm 0.11}$	0.6229 ± 0.13	0.4042 ± 0.13	0.5571 ± 0.14	0.3852 ± 0.19	0.6574 ± 0.12
$D_{prs} \downarrow$	Deep Learning	0.3805 ± 0.33	$\mathbf{0.3934 \pm 0.34}$	$\mathbf{0.2326 \pm 0.17}$	$\mathbf{0.1692 \pm 0.18}$	$\mathbf{0.3402 \pm 0.21}$	$\mathbf{0.1192 \pm 0.16}$
	Statistical	$\mathbf{0.3481 \pm 0.16}$	0.4679 ± 0.18	0.6802 ± 0.20	0.5376 ± 0.23	0.6361 ± 0.22	0.3199 ± 0.16
$F_1 - IoU \downarrow$	Deep Learning	**0.1021**	**0.0986**	**0.0961**	**0.0690**	**0.1164**	**0.0524**
	Statistical	0.1260	0.1348	0.1586	0.1461	0.1441	0.1279

Original | Ground truth | [MD] CNN | [MD] Statistical | [LD] CNN | [LD] Statistical

In this case the statistical approach has better F1, but worse IoU: the statistical approach picks more True Positives than the CNN.

In Medium on Dark case, the Dprs score of the statistical method is worse than in the case of Light on Dark, even if the F1 and IoU are better.
Specificity is driving the prediction away from the ideal ground truth, suggesting very few True Negatives.

[MD] Medium on Dark - Medium as training, Dark as testing

[LD] Light on Dark – Light as training, Dark as testing

Results: Inference time

The **thresholding** approach is 65x and 118x times faster than the statistical and the CNN, respectively, achieving 140 FPS.
It also has null standard deviation, which highlights the impartiality of the algorithm given different images.

	Inference time (seconds)
Deep Learning	0.826581 ± 0.043
Statistical	$\underline{0.457534 \pm 0.002}$
Thresholding	$\mathbf{0.007717 \pm 0.000}$

*Measured on a i7 4770k CPU and 16 GB of RAM

The first **14 ECU images**, with size of 352×288, have been used as testing dataset.
One image at a time has been processed by the methods and the resulting execution time has been saved.
The set of pictures has been processed 5 times and, each time, the averaged measurement time has been calculated.
Finally, the average values have been averaged into a single value and the standard deviation has been computed.

- Image loading into memory is excluded
- Image saving to disk is excluded
- The measurement starts when the algorithm starts
- Pre-processing and post-processing, if present, are included in the measured execution time

Conclusions and Future Work

- In-depth analysis of three main approaches of state-of-the-art
- Single and Cross datasets evaluation
- Single and Cross skin tones evaluation
- Inference Time evaluation

The generalization and semantic features extraction capabilities of **CNNs** have proven to be really powerful. **Thresholding** methods had the worst precision but proved to be really fast.
Involving **multiple metrics** have debunked over-optimistic results.

The future work could focus on **Transformers**, as they have proven to be really successful in Natural Language Processing [6] and are starting to gain traction in the image segmentation tasks [7,8], and **mobile Machine Learning**, which is becoming a solid platform for U-Nets [9].

References

[1] Ramirez, G. A., Fuentes, O., Crites Jr, S. L., Jimenez, M., & Ordonez, J. (2014). Color analysis of facial skin: Detection of emotional state. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 468-473).
[2] Do, T. T., Zhou, Y., Zheng, H., Cheung, N. M., & Koh, D. (2014, August). Early melanoma diagnosis with mobile imaging. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6752-6757). IEEE.
[3] Brancati, N., De Pietro, G., Frucci, M., & Gallo, L. (2017). Human skin detection through correlation rules between the YCb and YCr subspaces based on dynamic color clustering. *Computer Vision and Image Understanding*, *155*, 33-42.
[4] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.
[5] Tarasiewicz, T., Nalepa, J., & Kawulok, M. (2020, October). Skinny: A Lightweight U-net for Skin Detection and Segmentation. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 2386-2390). IEEE.
[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
[7] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.
[8] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv preprint arXiv:2105.05537*.
[9] Ignatov, A., Byeoung-su, K., Timofte, R., & Pouget, A. (2021). Fast camera image denoising on mobile gpus with deep learning, mobile ai 2021 challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2515-2524).

Thanks for the attention!