Fair comparison of skin detection approaches on publicly available datasets

Alessandra Lumini[a*] and Loris Nanni[b]

[a.] *DISI, Università di Bologna, Via dell'Università 50, 47521 Cesena, Italy.*
[b] *DEI - University of Padova, Via Gradenigo, 6 - 35131- Padova, Italy.*

ABSTRACT

Skin detection is the process of discriminating skin and non-skin regions in a digital image and it is widely used in several applications ranging from hand gesture analysis to tracking body parts and face detection. Skin detection is a challenging problem which has drawn extensive attention from the research community, nevertheless a fair comparison among approaches is very difficult due to the lack of a common benchmark and a unified testing protocol. In this work, we investigate the most recent researches in this field and we propose a fair comparison among approaches using several different datasets.

The major contributions of this work are an exhaustive literature review of skin color detection approaches, a framework to evaluate and combine different skin detector approaches, whose source code is made freely available for future research, and an extensive experimental comparison among several recent methods which have also been used to define an ensemble that works well in many different problems.

Experiments are carried out in 10 different datasets including more than 10000 labelled images: experimental results confirm that the ensemble here proposed obtains a very good performance with respect to other stand-alone approaches, without requiring ad hoc parameter tuning.

A MATLAB version of the framework for testing and ensemble proposed in this paper will be freely available from https://github.com/LorisNanni.

Keywords: skin classification, skin detection, skin segmentation, skin database.

1. Introduction

Skin texture and color are important signs that people use to understand variety of culture-related aspects about each other, as: health, ethnicity, age, beauty, wealth and so on. The presence of skin color in images and videos is a signal of the presence of humans in such media. Therefore in the last two decades extensive research has focused on skin detection in videos and images. Skin detection is the process of discriminating "skin" and "non-skin" regions in a digital image and consists in performing a binary classification of pixels and in executing a fine segmentation to define the boundaries of the skin regions.

Skin detection is used within many application domains: it is used as a preliminary step for face detection [1] and tracking [2], body tracking [3], hand gesture recognition [4], biometric authentication (i.e. palm print recognition) [5], objectionable content filtering [6], medical imaging.

A useful feature for the discrimination of skin and non-skin pixels is the pixel color; nevertheless, obtaining skin color consistency across variations in illumination, diverse ethnicity and different acquisition devices is a very challenging task. Moreover, skin detection, when used as preliminary step of other applications, is required to be computationally efficient, invariant against geometrics transformations, partial occlusions or changes of posture/facial expression, insensitive to complex or pseudo-skin background, robust against the quality of the acquisition device. The factor that worst influences skin detection is the color constancy problem [7]: i.e. the dependency of pixel intensity on both reflection and illumination which have a nonlinear and unpredictable behavior. To be effective when the illumination conditions vary rapidly, some skin detection approaches use image preprocessing strategies based on color constancy (i.e. a color correction method based on an estimate of the illuminant color) and/or dynamic adaptation techniques (i.e. the transformation of a skin-color model according to the changing illumination conditions). Static skin color approaches that rely on image preprocessing can only partially solve this problem and their performance strongly degrades in real-world applications. A possible solution is considering additional data acquired out the visual spectrum (i.e. infrared images [8] or spectral imaging [9]), however the use of such sensors is not appropriate for all applications and requires higher acquisition costs which limit their use to specific problems.

Skin detection is a challenging problem and has been extensive studied from the research community. Despite the large number of methods, there are only few surveys in this topic: the works in [10][11] are quite old and cover only the methods proposed before 2005, the surveys in [12][13][14] are more recent and contain a good investigation of methods, benchmarking datasets and performance related to a period of about two decades. Anyway in none of the above surveys is there a fair comparison among methods using the same testing protocols and datasets. The aim of the present work is not limited to survey the most recent research in this field (which is now enriched of methods based on deep learning [15][16][17][18]), but also, and above all, to propose a framework for a fair comparison among approaches.

In this research, a novel framework is proposed that integrates different skin color classification approaches and compare their

performance and their combination on several publicly available datasets. The major contributions of this research work are:

- An exhaustive literature review of skin color detection approaches with a detailed description of methods freely available.

- A framework to evaluate and combine different skin detector approaches. The source code of the framework and many of the tested methods will be made freely available for future research and comparisons. The system can be tuned according to the target application: on the basis of the application requirement, the acceptance threshold can be tuned to prune a large percentage of false accepts at a small cost of reduction in genuine accepts or vice versa a larger number of false accepts can be admitted to maximize the number of genuine accepts. The framework includes training and testing protocols for most used benchmark datasets in this field.

- A fair comparison among the most recent research and methods in the skin detection field, using the same testing protocols, benchmark datasets and performance indicators. A discussion about performance can help researchers and practitioners in evaluating the approaches most suited to their requirements according to computational complexity, memory requirements, detection rate and sensitivity.

The arrangement of this paper is as follows. In section 2 related works in skin detection are presented, including a discussion about taxonomy of existing approaches and a detailed description of the approaches tested in this work. In section 3 the evaluation problem is treated, the most known datasets used for performance evaluation are listed and commented, testing protocols and performance indicators used in our experiment are discussed. In section 4 the experiments conducted using the proposed framework are reported and discussed. Finally, section 5 includes the conclusions and some future research directions.

2. Skin detection approaches

Several skin detection methods are based on the assumption that skin color can be recognized from background colors according to some clustering rule in a specific color space. Even if this assumption can be valid in a constrained environment where both ethnicity of the people and background colors are known, it is a very challenging task in complex images captured under unconstrained conditions and when individuals show a large spectrum of human skin coloration [11]. There are a lot of challenging factors that influence the performance of a skin detector:

- Human characteristics as ethnicity and age: skin color ranges from white to dark brown among human racial groups, the transition from fresh skin to dry skin related to the age determines a strong variation of tones.

- Acquisition conditions: factors such as camera characteristics or illumination variations strongly influence the skin appearance. Generally, a variation in the illumination level or in the light source distribution determines the presence of shadows and changes in the color of the skin.

- Skin painting: the presence of makeup or tattoo coverage affects the appearance of skin.

- Complex background: sometimes the skin detector can be deceived by the presence of objects that have skin-like color in the background.

Skin detection approaches can be classified according to several taxonomy schemes which evaluate different aspects of the methods:

1. Considering the presence or not of preprocessing steps such as color correction and illumination cancelation [19] or dynamic adaptation [20] designed to reduce the effects of different acquisition conditions;

2. Considering the color space used for pixel classification [11]. Most of research papers dealing with the skin detection have to face the problem of selecting the most appropriate skin color model [14]. The performance of several different color models are compared in [11]: basic models (i.e. RGB, normalized RGB), perceptual models (i.e. HIS, HSV) perceptual uniform models (i.e. CIE-Lab, CIE-Luv) and orthogonal models (i.e. YCbCr, YIQ) with the finding that orthogonal models are characterized by a reduced redundancy/correlation among channels, therefore they are the most suited for skin detection.

3. Examining the problem formulation, which can be based on the segmentation of an image in the regions where human skin is present (segmentation based approaches) or on treating each pixel as skin or non-skin without considering its neighbors (pixel-based approaches). Despite the presence of a huge number of pixel-based approaches [11], the region-based skin color detection techniques are very few [21][22][23][24]. Some recent methods [16][17] based on convolutional neural networks can be included in this category.

4. Distinguish among methods for performing pixel classification (i.e. explicitly defined skin region, parametric approaches, non-parametric approaches) [25]. The first category, also named rule based, defines explicit rules to define the skin color (in an opportune color space), The other categories include machine learning approaches that make use of non-parametric or parametric approaches to estimate the color distribution from a training set (returning a lookup table or a parametric model, respectively).

5. Considering the type of classifier used for machine learning approaches: in [13] a taxonomy of 8 not exclusive groups is proposed which extends the simple division in parametric and non-parametric approaches. *Statistical methods* include parametric methods based on Bayesian rule o Mixture models [26]. *Neural networks models* [27][28] are used for the segmentation of color images based on both color and texture information. *Diffusion based methods* [29][30] extend analysis on neighboring pixels to improve the classification performance: after an initial pixel based extraction a seed growing method is applied to include neighborhood in the skin region. *Adaptive methods* [31] are based on the tuning of models in order to adapt to specific conditions (i.e. lighting, skin tone, background); the model calibration often grants performance advantages, but at the cost of increased computation time. *Hyperspectral models* [32] are based on acquisition devices with hyperspectral capabilities: despite the advantage due to the presence of spectral information, these approaches will be not considered in this work since they are only applicable to specifically collected datasets. *SVM based systems* are parametric models based on SVM classifier: this class also overlaps adaptive methods if the SVM classifier is trained using active learning [33]. *Mixture techniques* are the methods based on the mixture of different machine learning approaches [34][35].

In the last few years the research in skin detection has taken two main directions, according to the following consideration: even if in the most general applications nothing should be assumed about the background and the acquisition conditions, in many applications the difference between skin and background is large, the acquisition conditions are controlled and the skin region is quite easy to detect. For example, in many gesture recognition applications the hand images are acquired using flatbed scanners and have dark unsaturated backgrounds [36]. For this reason, in addition to many approaches that adopt sophisticated and computationally *expensive techniques*, several *simple rule-based methods* have been proposed, which are preferred in some applications since they are easier to understand, implement and

reuse, more efficient, while at the same time, adequately effective. Usually *simple rule-based methods* are not even tested on benchmarks for pure skin detection but as a step of a more complex task (i.e. face detection, hand gesture recognition). A recent example of method belonging to this class is the work in [36], which performs a study on different color models, drawing the conclusion that there are no apparent advantages of using a perceptual uniform color space: therefore they propose an approach based on a simple RGB lookup table. The work in [37] proposes a dynamic but very straightforward method based on parametric modeling the Cr or Cb channel from YCbCr color space through a single Gaussian: the final approach requires a limited amount of storage space, is fast and can be trained using a small training set with small time delay. In [38] a simple skin detector based on RGB histogram thresholding is employed as a preliminary fast step for motion-based skin region of interest detection. A dynamic skin color model method based on space switching is proposed in [39], where in order to handle with natural changes in illumination, a system with three robust algorithms has been built, based on different color spaces which are switched according to the statistical mean of the skin pixels in the image. In [40] a new 3D hybrid color space named SKN is proposed, obtained by using Principal Component Analysis and a Genetic Algorithm to discover the optimal representation of skin color over 17 color spaces, then a pixel-wise skin classification is performed employing a general purpose trained classifier (i.e. Naïve Bayes, Random Forest, Multilayer Perceptron and Support Vector Machines).

The first class of methods, based on sophisticated and computationally *expensive techniques*, include recently proposed approaches based on deep learning [15][16][17][18]. Convolutional neural networks have recently achieved remarkable results for a variety of computer vision tasks, including several applications based on pixel-wise prediction (i.e. scene labelling, semantic image segmentation). In [15] a patch-wise skin segmentation approach is described based on deep neural networks, which uses image patches as processing units instead of pixels; a dataset of image patches have been appositely collected for training purposes and the trained deep skin models are integrated into a sliding window framework to detect skin regions of the human body parts achieving competitive performance on pixel-wise skin detection. In [16] an integration of some recurrent neural networks layers into the fully convolutional neural networks is proposed as a solution to develop an end-to-end network for human skin detection. In [17] the authors propose a inception-like convolutional network-in-network structure, which consists of convolution and rectified linear unit layers only (without pooling and subsampling layers), while in [18] the authors make experiments on many CNN structures to determine the best one for skin detection. All the proposed architectures report comparable or better performance than other the state-of-the-art methods; anyway, a fair comparison with traditional existing approaches is always difficult due to the different testing protocols. Since none of the authors above has released a model to perform comparisons, in this work we train and use two of the most recently proposed models for image segmentation: SegNet [41] and U-Net [42]. SegNet is a convolutional encoder-decoder net: it uses VGG16 or VGG19 with only forward connections and non trainable layers as encoder, then for each of the encoder layers there is a corresponding decoder which upsamples the feature map using max-pooling indices. The U-Net architecture is a combination of convolutions and transposed convolutions layers used downsampling and then upsampling the images spatially. The -Net architecture has reported impressive results in medical image segmentation.

In this work, we evaluate and combine several skin detectors with the aim of comparing their performance and we propose an ensemble able of maximizing their classification performance. The base approaches used to create the ensemble have been selected according to their effectiveness, their availability to research scopes and their efficiency:

- *GMM* [26] is a simple and efficient Gaussian mixture model for skin detection trained to classify non-skin vs. skin pixels in the RGB space.

- *Bayes* [26] is a fast and effective method based on a Bayesian classifier. In this work we used a classifier trained in the RGB color space using 2000 images from the ECU data set.

- *SPL*(τ) [43][1] is a pixel-based skin detection approach which determines skin probability in the RGB domain using a look up table (LUT). For a testing image the probability of being skin for each pixel x is calculated, then a threshold τ is applied to decide *skin/non-skin*.

- *Cheddad*(τ) [44] is a pixel-based and real-time approach which reduces the RGB color space to a 1D space derived from differentiating the grayscale map and the non-red encoded grayscale version. The classification is performed using a skin probability which delimits the lower and upper bounds of the skin cluster and the final decision depends on a classification threshold τ;

- *Chen* [45] is a statistical skin color model, which is specifically suited to be implemented on hardware. The 3D skin cube is represented as three 2D sides calculated as the difference of two color channels: sR=R-G, sG=G-B, sB=R-B. The skin cluster region is delineated in the transformed space, fixing the boundaries to the following ranges: $-142 < sR < 18$, $-48 < sG < 92$, and $-32 < sB < 192$.

- *SA1*(τ) [46] is a method for skin detection based on spatial analysis. Starting from a skin probability map obtained using a color pixel-wise detector, the starting step to the spatial analysis is the proper selection of the high probability pixels as "skin seeds". The second step consists in finding the shortest routes from each seed to every single pixel, in order to propagate the "skinness". All the pixels not adjoined during the propagation process are labeled as non-skin. The performance depends on the threshold τ used to classify.

- *SA2*(τ) [47] is another recent method based on spatial analysis which uses both color and textural features to determine the presence of skin. The basic idea is to extract the textural features from the skin probability maps rather than from the luminance channel: therefore simple textural statistics are computed from each pixel's neighborhood in the probability map using kernels of different sizes (i.e. the median, the minimal values, the difference between the maximum and minimum and the standard deviation). Then skin and non-skin pixels are transformed into two classes of feature vectors whose size is reduced by Linear Discriminant Analysis (LDA) to increase their discriminating power. Finally the spatial analysis method proposed in [46] and described above is used for seed extraction and propagation using the distance transform. The classification depends on a threshold τ used to classify pixels in the distance domain.

- *SA3*(τ) [31] is a self-adaptive method that consists in combining a local skin color model created using a probability map and spatial analysis to fix the boundaries of the skin regions. It is an evolution of the method proposed in [47] based on spatial analysis, skin color model adaptation and textural features. The main difference from the method in [47] is a new technique for extracting adaptive seeds, based on the analysis of the skin probability map calculated from the input color image. The performance depends on a classification threshold τ.

---

[1] http://clickdamage.com/sourcecode/index.php

- *DYC* [48][2] is a skin detection approach which works in the YCbCr color space and takes into account the lighting conditions. The method is based on the dynamic generation of the skin cluster range both in the YCb and YCr subspaces of YCbCr and on the definition of correlation rules between the skin color clusters.

- SegNet has been modeled starting from a pre-trained VGG19 network and then fine-tuned using the first 2000 images of the ECU dataset and the following options: batch size 32, learning rate 0.001, max epoch 45 (data augmentation, i.e. flip and small rescaling, has been used only in the first 15 epoch). Moreover we use inverse class frequency weighting in order to deal with imbalanced classes: the frequencies of "skin" and 'non-skin' labels over the training data are used as 'ClassWeights' in final pixelClassificationLayer.

- U-Net has been trained from scratch using the first 2000 images of the ECU dataset and the following options: batch size 32, max epoch 45 (data augmentation has been used only in the first 15 epoch), learning rate 0.0001 only in the first 15 epoch then 0.001, class weighting.

A rough classification of the methods tested in this work according to the classification criteria surveyed in section 2 is reported in Table 1.

3. Skin detection evaluation: Datasets and performance indicators

To assist research in the area of skin detection, there are some well-known color image datasets provided with ground truth. The use of a standard and representative benchmark is essential to execute a fair empirical evaluation of skin detection techniques.

Table 1. Rough classification of the tested approaches

	GMM	Bayes	SPL	Cheddad	Chen	SA1	SA2	SA3	DYC	SegNet	U-Net
Preprocessing steps											
None	x	x	*x*	x	x					x	x
Dynamic adaptation						x	x	x	x		
Color space											
Basic color spaces	x	x	x							x	x
Perceptual color spaces						x	x	x			
Orthogonal color spaces									X		
Other (e.g. Color ratio)				x	x						
Problem formulation											
Segmentation based						x	x	x		x	x
Pixel based	x	x	x	x	x				x		
Type of pixel classification											
Rule based				x	x				x		
Machine learning: parametric	x	x									
Machine learning: non-parametric			x								
Type of classifier											
Statistical		x	x								
Mixture techniques	x										
Adaptive methods						x	x	x			
CNN										x	x

Table 2. Some of the most used datasets per skin detection

Name	Ref	Images	Ground truth	Download	Year
Compaq	[26]	4675	Semi-supervised	ask to the authors	2002
TDSD	[49]	555	Imprecise	http://lbmedia.ece.ucsb.edu/research/skin/skin.htm	2004
UChile	[50]	103	Medium Precision	http://agami.die.uchile.cl/skindiff/	2004
ECU	[51]	4000	Precise	http://www.uow.edu.au/~phung/download.html (currently not available)	2005
Schmugge	[52]	845	Precise (3 classes)	https://www.researchgate.net/publication/257620282_skin_image_Data_set_with_ground_truth	2007
Feeval	[53]	8991	Low quality, imprecise	http://www.feeval.org/Data-sets/Skin_Colors.html	2009
MCG	[54]	1000	Imprecise	http://mcg.ict.ac.cn/result_data_02mcg_skin.html (ask to authors)	2011
Pratheepan	[55]	78	Precise	http://web.fsktm.um.edu.my/~cschan/downloads_skin_dataset.html	2012
VDM	[56]	285	Precise	http://www-vpu.eps.uam.es/publications/SkinDetDM/	2013
SFA	[57]	1118	Medium Precision	http://www1.sel.eesc.usp.br/sfa/	2013
HGR	[31]	1558	Precise	http://sun.aei.polsl.pl/~mkawulok/gestures/	2014
SDD	[58]	21000	Precise	Not available yet	2015

---

[2] https://github.com/nadiabrancati/skin_detection/

3.1. Datasets

In Table 2 some of the most used datasets are summarized and in this section a brief description of each of them is given.

Compaq [26] is one of the first and most used large skin dataset, it consists of images collected by crawling Web: 9731 images containing skin pixels (but only 4675 skin images have been segmented and included in the ground truth) and 8965 images with no skin pixels. This dataset have been extensity used for testing and comparing methods, but without using a standard testing protocol, therefore comparisons using this dataset are not always fair. Moreover, the ground truth for this dataset has been obtained using an automatic software tool leading to imprecise results.

TDSD [49] is an old dataset containing 555 images with very imprecise annotations (automatic labeling).

The **UChile** [50] dbskin2 complete set includes 103 images acquired in different lighting conditions and with complex background. The ground truth has been manually annotated with medium precision (in some images the boundaries between skin and non-skin pixels are not marked precisely).

ECU [51] skin and face datasets are a collection of about 4000 color images annotated with relatively accurate ground truth. ECU dataset is quite challenging, since it includes diversity in terms of the lighting conditions, background scenes, and skin types.

The skin dataset named **Schmugge** [52] is a collection of 845 images taken from different face datasets (i.e. the UOPB dataset, the AR face dataset, and University of Chile database). The ground truth is very accurate since all images are labelled in 3 classes: skin/ not-skin/don't care.

Feeval [53] is a dataset based on 8991 frames extracted from 25 online videos of low quality. The ground truth is not precisely annotated. Here, the performance indicator is calculated averaging the performance of each video by the number of frames (therefore considering each video as a single image in the performance evaluation).

MCG skin database [29] contains 1000 images selected from internet in order to include confusing backgrounds, variable ambient lights and diversity of human races. The ground-truth is obtained through manually labeling, but it is not very precise since eyes, eyebrows, and even bracelets are sometimes marked as skin.

VMD [56] is obtained selecting 285 images from several public datasets for human activity recognition (i.e. LIRIS, EDds, UT, AMI and SSG). The dataset is already divided in subsets for training and testing, anyway in this work we used all the images for testing. The images cover a wide range of illumination levels and situations.

SFA dataset [57] includes images from FERET (876 images) and AR (242 images) face databases manually labelled (with medium precision). SFA includes an internal organization in folders in order to separate 1118 original images (ORI), 1118 ground truths (GT) masks, 3354 samples of skin (SKIN) and 5590 samples of nonskin (NS) ranging differently from 1 to 35×35 dimensions. We have used ORI/GT for assessing the performance.

Pratheepan [55] is a small dataset which includes 78 images downloaded randomly from Google; the dataset is divided in two subsets: FacePhoto including 32 single subject images with simple background and FamilyPhoto including 46 images with multiple subjects and complex background.

HGR [31] is a dataset for gesture recognition which contains also ground truth binary skin presence masks; the dataset includes 1558 images representing gestures from Polish and American sign language with controlled and uncontrolled background divided in 3 subsets (HGR1, HGR2A, HGR2B). In our tests the size of the images of HGR2A and HGR2B has been reduced of a factor 0.3.

SDD [58] is a dataset of 21000 images acquired using different imaging devices, in a variety of illumination conditions and including different skin colors from people all around the world. The dataset is composed from some images extracted from videos and others belonging to popular face datasets. Images are provided in four sets: a set including mainly single face images, made for training purposes, and other three sets to be used for testing.

FvNF [59] (Face vs. NonFace) is not a real skin dataset, it is composed by 800 face and 770 non-face images, extracted from Caltech dataset [60]. This dataset has been collected and used in [59] to evaluate the capability of a skin detector method to detect the presence of a face, based on the number of pixels classified as skin. The average precision (*AP*) is used as performance indicator, $AP \in [0,100]$. The *AP* summarizes the shape of the precision/recall curve, since it is the area under the precision-recall curve.

3.2. Performance indicators

Skin detection is a two-class classification problem, therefore standard measures for general classification problems [61] can be used for performance evaluation: including Accuracy, Precision, Recall, F_1 measure, Kappa, ROC-Curve, Area Under Curve, and others. Anyway, due to the particular nature of this problem which is based on pixel-level classification and on unbalanced distribution, the following measures are usually considered for performance evaluation: the confusion matrix, the F_1 measure, the True Positive Rate (*TPR*) and the False Positive Rate (*FPR*).

The *confusion matrix* is obtained comparing results with the ground-truth data to determine the number of true negatives (*tn*), false negatives (*fn*), true positives (*tp*) and false positives (*fp*). Several useful measures can be obtained from the confusion matrix, including *precision=tp*/(*tp+fp*), that is the percentage of correctly classified pixels out of all the pixels classified as skin and *recall =tp*/(*fn+tp*), that is the percentage of the ground-truth skin pixels correctly classified as skin.

The *F_1 measure* is the harmonic mean of precision and recall and it is calculated according to the following formula $F_1 = 2tp/(2tp + fn + fp)$, $F_1 \in [0,1]$. According to other works in the literature F_1 is averaged at pixel level not at image level; in such a way the final indicator is not dependent on the image size in the different databases. Besides F_1 several papers use the *True Positive Rate* (*TPR= recall=tp*/(*fn+tp*)) and the *False Positive Rate* (*FPR= fp*/(*fp+tn*)).

Even if the *F_1 measure* is able to evaluate algorithms only at a fixed operating threshold, and therefore it is a worse way to quantify the accuracy of an algorithm evaluation with respect to ROC curve or precision-recall curve, we use F_1 in this work since it is widely used in the literature for skin classification and allow a better comparison. Moreover, several methods evaluated here use a fixed threshold and do not allow tuning for true positives and false positives rates.

4. A fair experimental comparison

A fair comparison among different approaches is very difficult due to the lack of a universal standard in evaluation: most of published works are tested on self-collected datasets which often are not available for further comparison; in many cases the testing protocol is not clearly explained, many datasets are not of high quality and the precision of the ground truth is questionable since sometimes lips, mouth, rings and bracelets have been labelled as skin. In this section, we carry out a comparison of some well-known approaches whose source code is freely available. In order to accomplish a fair comparison we do not perform re-training of methods in each dataset, conversely, we use the knowledge provided by the original authors, therefore all the datasets have been used only for testing purposes. When the performance depends on a classification threshold, we tested several values in order to select the one that performs best in all the datasets.

In Table 3 we report the performance in 10 datasets of the nine stand-alone approaches described in section 2 and three ensembles combined by the vote rule. Both the stand-alone approaches and the ensembles have been tested using different thresholds (only the best one is reported in Table 3 for sake of space, the complete version of the table is included as supplementary material):

- *GMM* [26], *Bayes* [26] ($\tau\in[50,70,90,110,140]$), *SPL* [43] ($\tau\in[-2.5,-2,-1.5,-1,-0.5]$), *Cheddad* [44], *Chen* [45], *SA1* [46] ($\tau\in[100,150,175,200,225]$), *SA2* [47] ($\tau\in[30,40,50,85,120]$), *SA3* [31] ($\tau\in[25,50,75,100,125]$), *DYC* [48]

- *SegNet, U-Net* are the method based on CNN trained on an external dataset according to the procedure explained in section 2.

- *Vote1*, *Vote2* and *Vote3* are the ensembles proposed in this work and based on the weighed fusion of the following methods using the vote rule: $w_1\times SA1(175) + w_2\times SA2(50) + w_3\times SA3(50) + w_4\times Cheddad(125) + w_5\times DYC + w_6\times Bayes(110) + w_7\times SegNet + w_8\times U\text{-}Net > (w_1 + w_2 + w_3 + w_4 + w_5 + w_6 + w_7 + w_8)/w_\tau$, where $\mathbf{w}=(w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8)$ is the weight vector and $w_\tau\in[1.25,1.5,1.75,2]$ is a threshold parameter used to tune the sensitivity of the method. The weight arrays used for the three ensemble are, respectively:

  w1=(0.5, 1.5, 1, 1.5, 0.5, 1, 0, 0),
  w2=(0.5, 1.5, 1, 1.5, 0, 1, 5.5, 0),
  w3=(0.5, 1.5, 1, 1.5, 0, 1, 5.5, 2.75);

therefore the first ensemble (*Vote1*) does not include any CNN classifier, *Vote2* includes *SegNet* (weighted as the sum of the other classifiers) and *Vote3* also includes *U-Net* (weighed half of SegNet). In our experiments we have tested different rules for the fusion of classifiers, finding that the vote rule is the best suited for this particular problem.

For each dataset the best result is highlighted and in the last column the rank (calculated as the rank of the average rank) is reported. Maybe some approaches can be advantaged in some datasets, in case that they used some images of the dataset for training (i.e. LUT creation), anyway considering the results on all the 10 datasets, the comparison is quite fair.

From the results in Table 3 it is clear that no approach overcomes the others in all the problems, on the contrary it is very difficult to find a solution that works well in all the datasets without retraining and performing a parameter tuning. The design of an ensemble method able to exploit the good performance of each of its composing approaches is a feasible solution to this problem, indeed our proposed ensembles based on the vote rule grant good performance on average in almost all the datasets and reach the first positions in the rank column. This is a valuable result considering that our experiments compare 14 approaches on 10 different datasets.

Table 3. Performance in 10 different datasets

Perf. Indicator		Dataset									AP	Rank
		F_1 measure										
Method	τ/w_τ	Feeval	Pratheepan	MCG	UChile	Compaq	SFA	HGR	Schmugge	VMD	FnF	
GMM		0.513	0.581	0.688	0.615	0.600	0.789	0.658	0.595	0.130	91.590	12
Bayes	110	0.565	0.631	0.694	0.661	0.599	0.760	0.871	0.569	0.252	92.380	13
SPL	-2	0.544	0.551	0.621	0.568	0.494	0.700	0.845	0.490	0.321	89.380	7
Cheddad	125	0.596	0.597	0.667	0.649	0.588	0.683	0.767	0.571	0.261	89.200	10
Chen		0.287	0.540	0.656	0.598	0.549	0.791	0.732	0.571	0.165	87.200	14
SA1	175	0.558	0.613	0.664	0.567	0.593	0.788	0.768	0.482	0.199	86.770	11
SA2	50	0.515	0.693	0.755	0.663	0.645	0.771	0.806	0.594	0.156	91.910	8
SA3	50	0.539	0.709	0.762	0.625	0.647	0.863	0.877	0.586	0.147	89.800	6
DYC		0.588	0.599	0.680	0.663	0.618	0.569	0.616	0.613	0.275	88.710	9
SegNet		0.717	0.730	0.813	0.802	0.737	0.889	0.869	0.708	0.328	**99.978**	3
U-Net		0.576	0.787	0.779	0.713	0.686	0.848	0.836	0.671	0.332	98.848	4
Vote1	1.5	0.594	0.717	0.754	0.670	0.666	0.737	0.849	0.625	0.269	92.980	5
Vote2	1.75	**0.736**	0.811	0.816	0.81	0.772	0.854	0.949	0.700	**0.481**	99.780	2
Vote3	1.75	0.733	**0.812**	**0.841**	**0.829**	**0.773**	**0.902**	**0.950**	**0.714**	0.423	99.800	**1**

From a comparison among stand-alone approaches in table 3, the following conclusions can be drawn:

- A very noticeable aspect is the great performance improving given by the use of deep learning: the two CNNs are the best stand-alone approaches in almost all the datasets and perform better than the fusion of all handcrafted methods (*Vote1*). For this reason they are weighed more than other approaches in our ensembles *Vote2* and *Vote3* (*SegNet* is weighted as the sum of the handcrafted classifiers and *U-Net* is weighted half of *SegNet*).

- In particular SegNet is the best stand-alone method. The great advantage with respect to U-Net is that the training of SegNet has been performed starting from a pre-trained model (VGG19), while U-Net has been trained starting from random weights. This is a not negligible aspect considering that our training set contains only 2,000 images. The performance of both the networks can be certainly boost by a further training on a more specific dataset: i.e. classification in HGR may be improved by training using hand recognition images.

- The ensembles *Vote2* and *Vote3* have similar performance, even if Vote3 outperforms Vote2 (p-value 0.2305 using Wilcoxon signed rank test [62]). Both the ensembles outperforms all the other approaches (p-value di 0.05).

- If we consider only the handcrafted methods, there is a noticeable difference among performance of different approaches in each dataset. The best stand-alone method is different for each dataset, anyway *SA2* and *SA3* work better than others do, this is the reason why we gave them the higher weights in our vote rule. In particular, *SA3* reaches the best performance in three datasets: *MCG*, *SFA*, *HGR*. In our opinion, it is a very valuable method in particular for problems where the background is quite simple (e.g. gesture recognition).

- The method *Bayes* perform very well (it is ranked third among handcrafted stand-alone methods) even though it is a very simple approach with low computational requirements.

- The statistical color model proposed by *Chen* is very efficient but at the expense of performance. Its low performance (last position in the rank) is certainly related to the absence of a tuning parameter.

- The method *SPL* performs very well in *VMD* but very poorly in *Compaq*. SPL, which is based on lookup tables, is a method suited for simple tasks, as gesture recognition, where the image are taken in fixed acquisition condition and without complex background. This is the reason of its poor performance in *Compaq* which is a very difficult dataset.

For many approaches, the selection of an appropriate threshold is crucial for performance: the most appropriate choice depends of the specific application, since a strict value decreases the number of false positives despite of the true positive rate, vice versa a low threshold increases the true positive rate but also the number of false positives.

The ensemble that we propose as the best solution is *Vote3*(1.5) which include both the CNN models; besides F_1 measure we also report *TPR* and *FPR* as performance indicators for *Vote3*(1.5) in the MGC and UChile datasets, for an easier comparison with several methods in the literature that use these two indicators in these datasets:

- *MGC* (*TPR*=0.8077, *FPR*=0.0381),

- *UChile* (*TPR*=0.7984, *FPR*=0.0266).

Visual comparisons are also presented in Fig1, 2 and 3, which show the results of the tested methods on three different samples taken in various illumination condition and with different background. The three images are taken from the Schmugge dataset: (from left top to right bottom) input, ground truth, GMM, Bayes, SPL, Cheddad, Chen, SA1, SA2, SA3, DYC, SegNet, U-Net, Vote3.

Figure 1. Schmugge005.jpg: (from left top to right bottom) input, ground truth, GMM, Bayes, SPL, Cheddad, Chen, SA1, SA2, SA3, DYC, SegNet, U-Net, Vote3.
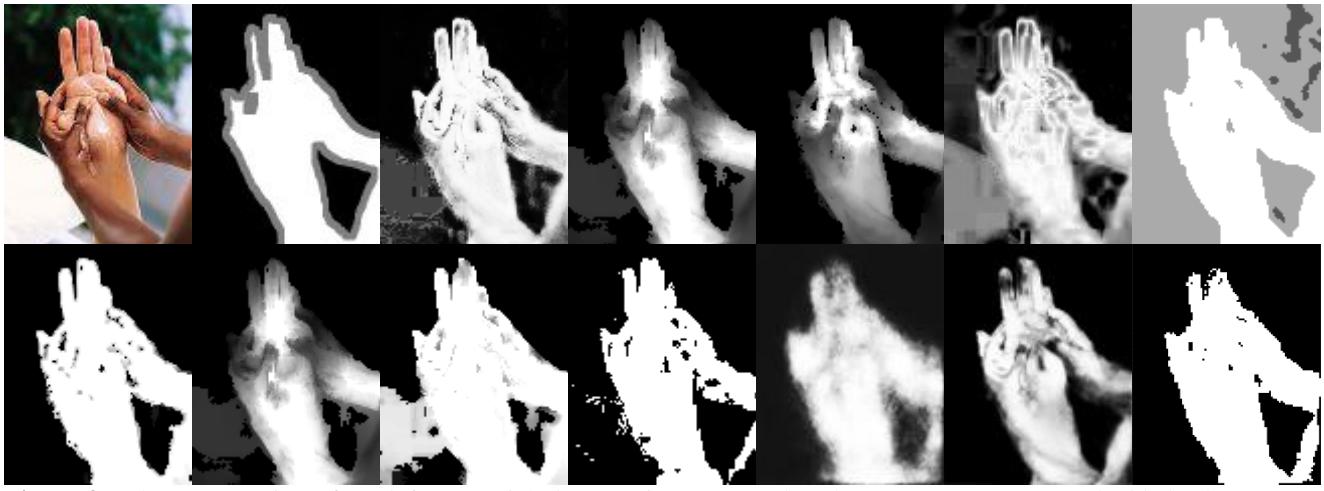
Figure 2. Schmugge367.jpg: (from left top to right bottom) input, ground truth, GMM, Bayes, SPL, Cheddad, Chen, SA1, SA2, SA3, DYC, SegNet, U-Net, Vote3.
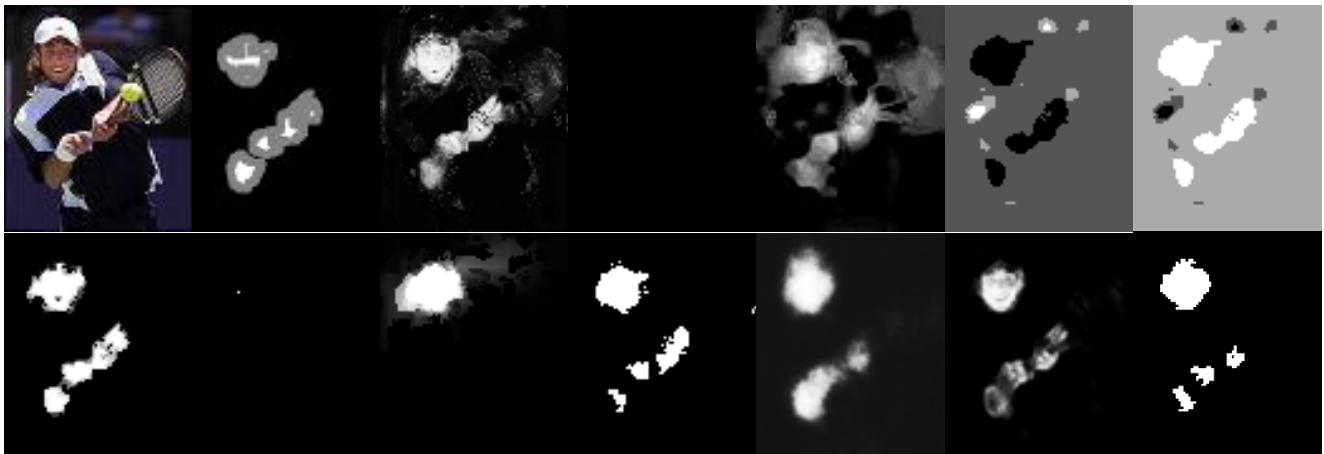

Figure 3. Schmugge845.jpg: (from left top to right bottom) input, ground truth, GMM, Bayes, SPL, Cheddad, Chen, SA1, SA2, SA3, DYC, SegNet, U-Net, Vote3.

5. Conclusion and future research directions

In this work a new framework to evaluate and combine different skin detector approaches is presented and an extensive evaluation of several approaches is carried out on 10 different datasets including more than 10000 labelled images. A survey of most recent existing approaches is carried out and a new ensemble based on the combination of eight methods (including 2 CNNs) is proposed and evaluated. Experimental results confirm that our proposed ensemble obtains a very good performance with respect to other stand-alone approaches. The proposed ensemble can be used in a wide range of skin detection applications such as body tracking, face detection, detection of objectionable content and hand gesture analysis. The performance obtained on 10 different datasets confirms that the proposed ensemble is well suited for detecting skin in a wide range of images without requiring re-training and ad hoc parameter optimization. Anyway, considering the impact on the performance of the "deep learning revolution" we cannot avoid suggesting to start from a trained CNN and fine-tuning on the particular problem to have very valuable performance.

In conclusion, we show that skin detection is a very difficult problem, which can be hardly solved by a stand-alone approach. The performance of any stand-alone method for skin detection depends on many factors, such as the color space used, the parameters used, the nature of data, the image characteristics, the shape of the distribution, the size of samples for training, the presence of data noise, etc. Our experiments show that our ensemble suffers less of these problems, since it is able to exploit the advantages of each composing method and it is a feasible solution for problems where training is impracticable.

Future work for designing an effective skin-color detector can be done by adding a preprocessing module able to deal with the color constancy problem in order to improve the robustness to illumination changes. Another idea to improve the precision of the detection is the use of morphological operators as a post-processing step. As future work for improve the performance of the ensemble we will test some adaptive approaches for selecting weights based on image content/quality and we will add other CNN models to the ensemble.

References

[1]   R.L. Hsu, M. Abdel-Mottaleb, A.K. Jain, Face detection in color images, IEEE Trans. Pattern Anal. Mach. Intell. (2002). doi:10.1109/34.1000242.

[2]   M. De-La-Torre, E. Granger, P.V.W. Radtke, R. Sabourin, D.O. Gorodnichy, Partially-supervised learning from facial trajectories for face recognition in video surveillance, Inf. Fusion. (2015). doi:10.1016/j.inffus.2014.05.006.

[3]   A. Argyros, M. Lourakis, Real-time tracking of multiple skin-colored objects with a possibly moving camera, Eccv 2004. vol3 (2004) 368–379. doi:10.1007/978-3-540-24672-5_29.

[4]   K. Roy, A. Mohanty, R.R. Sahay, Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation, in: IEEE Int. Conf. Comput. Vis. Work., 2017: pp. 640–649. doi:10.1109/ICCVW.2017.81.

[5]   H. Sang, Y. Ma, J. Huang, Robust Palmprint Recognition Base on Touch-Less Color Palmprint Images Acquired, J. Signal Inf. Process. 04 (2013) 134–139. doi:10.4236/jsip.2013.42019.

[6] J.-S. Lee, Y.-M. Kuo, P.-C. Chung, E.-L. Chen, Naked image detection based on adaptive and extensible skin color model, Pattern Recognit. 40 (2007) 2261–2270. doi:10.1016/j.patcog.2006.11.016.

[7] R. Khan, A. Hanbury, J. St??ttinger, A. Bais, Color based skin classification, Pattern Recognit. Lett. 33 (2012) 157–163. doi:10.1016/j.patrec.2011.09.032.

[8] S.G. Kong, J. Heo, B.R. Abidi, J. Paik, M.A. Abidi, Recent advances in visual and infrared face recognition - A review, Comput. Vis. Image Underst. 97 (2005) 103–135. doi:10.1016/j.cviu.2004.04.001.

[9] G. Healey, M. Prasad, B. Tromberg, Face recognition in hyperspectral images, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2003) 1552–1560. doi:10.1109/TPAMI.2003.1251148.

[10] C. Prema, M. Manimegalai, Survey on Skin Tone Detection using Color Spaces, Int. J. Appl. Inf. Syst. 2 (2012) 18–26.

[11] P. Kakumanu, S. Makrogiannis, N. Bourbakis, A survey of skin-color modeling and detection methods, Pattern Recognit. 40 (2007) 1106–1122. doi:10.1016/j.patcog.2006.06.010.

[12] W. Chen, K. Wang, H. Jiang, M. Li, Skin color modeling for face detection and segmentation: a review and a new approach, Multimed. Tools Appl. 75 (2016) 839–862.

[13] M.R. Mahmoodi, S.M. Sayedi, A Comprehensive Survey on Human Skin Detection, Int. J. Image, Graph. Signal Process. 8 (2016) 1–35. doi:10.5815/ijigsp.2016.05.01.

[14] S. Naji, H.A. Jalab, S.A. Kareem, A survey on skin detection in colored images, Artif. Intell. Rev. (2018). doi:10.1007/s10462-018-9664-9.

[15] T. Xu, Z. Zhang, Y. Wang, Patch-wise skin segmentation of human body parts via deep neural networks, J. Electron. Imaging. 24 (2015) 043009. doi:10.1117/1.JEI.24.4.043009.

[16] H. Zuo, H. Fan, E. Blasch, H. Ling, Combining Convolutional and Recurrent Neural Networks for Human Skin Detection, IEEE Signal Process. Lett. 24 (2017) 289–293.

[17] Y. Kim, I. Hwang, N.I. Cho, Convolutional neural networks and training strategies for skin detection, in: IEEE Int. Conf. Image Process., 2017: pp. 3919–3923. doi:10.1109/ICIP.2017.8297017.

[18] C. Ma, H. Shih, Human Skin Segmentation Using Fully Convolutional Neural Networks, in: IEEE 7th Glob. Conf. Consum. Electron., 2018: pp. 168–170. doi:10.1109/GCCE.2018.8574747.

[19] B.D. Zarit, B.J. Super, F.K.H. Quek, Comparison of five color models in skin pixel classification, Proc. Int. Work. Recognition, Anal. Track. Faces Gestures Real-Time Syst. Conjunction with ICCV'99 (Cat. No.PR00378). (1999) 58–63. doi:10.1109/RATFG.1999.799224.

[20] N.B. Ibrahim, M.M. Selim, H.H. Zayed, A Dynamic Skin Detector Based on Face Skin Tone Color, in: 8th Int. Conf. Informatics Syst., 2012: pp. 1–5.

[21] R.P.K. Poudel, J.J. Zhang, D. Liu, H. Nait-Charif, Skin Color Detection Using Region-Based Approach, Int. J. Image Process. 7 (2013) 385.

[22] W.C. Chen, M.S. Wang, Region-based and content adaptive skin detection in color images, Int. J. Pattern Recognit. Artif. Intell. 21 (2007) 831–853. doi:10.1142/s0218001407005715.

[23] H. Kruppa, M.A. Bauer, B. Schiele, Skin Patch Detection in Real-World Images, in: Annu. Symp. Pattern Recognit. DAGM, 2002: p. 109f. doi:10.1007/3-540-45783-6.

[24] N. Sebe, I. Cohen, T.S. Huang, T. Gevers, Skin detection: a Bayesian network approach, Proc. 17th Int. Conf. Pattern Recognit. 2 (2004) 2–5. doi:10.1109/ICPR.2004.1334405.

[25] A. Kumar, S. Malhotra, Pixel-Based Skin Color Classifier: A Review, Int. J. Signal Process. Image Process. Pattern Recognit. 8 (2015) 283–290.

[26] M.J. Jones, J.M. Rehg, Statistical color models with application to skin detection, Int. J. Comput. Vis. 46 (2002) 81–96. doi:10.1023/A:1013200319198.

[27] QingXiang Wu, R. Cai, L. Fan, C. Ruan, G. Leng, Skin detection using color processing mechanism inspired by the visual system, in: IET Conf. Image Process., 2012: pp. B10–B10. doi:10.1049/cp.2012.0459.

[28] L. Chen, J. Zhou, Z. Liu, W. Chen, G. Xiong, A skin detector based on neural network, in: Commun. Circuits Syst. West Sino Expo. IEEE 2002 Int. Conf., 2002: pp. 615–619 vol.1. doi:10.1109/ICCCAS.2002.1180694.

[29] M.R. Mahmoodi, S.M. Sayedi, Leveraging spatial analysis on homogonous regions of color images for skin classification, in: 4th Int. EConference Comput. Knowl. Eng., 2014: pp. 209–214.

[30] R. Nidhu, M.G. Thomas, Real Time Segmentation Algorithm for Complex Outdoor Conditions, Int. J. Sci. Technoledge. 2 (2014) 71.

[31] M. Kawulok, J. Kawulok, J. Nalepa, B. Smolka, Self-adaptive algorithm for segmenting skin regions, EURASIP J. Adv. Signal Process. (2014) 1–22. doi:10.1186/1687-6180-2014-170.

[32] A.S. Nunez, M.J. Mendenhall, Detection of Human Skin in Near Infrared Hyperspectral Imagery, Int. Geosci. Remote Sens. Symp. 2 (2008) 621–624. doi:10.1109/IGARSS.2008.4779069.

[33] J. Han, G.M. Award, A. Sutherland, H. Wu, Automatic skin segmentation for gesture recognition combining region and support vector machine active learning, in: Proc. 7th Int. Conf. Autom. Face Gesture Recognit., 2006: pp. 237–242. doi:10.1109/FGR.2006.27.

[34] Z. Jiang, M. Yao, W. Jiang, Skin Detection Using Color, Texture and Space Information, in: Fourth Int. Conf. Fuzzy Syst. Knowl. Discov., 2007: pp. 366–370. doi:10.1109/FSKD.2007.518.

[35] H.K. Al-Mohair, J. Mohamad Saleh, S.A. Suandi, Hybrid Human Skin Detection Using Neural Network and K-Means Clustering Technique, Appl. Soft Comput. 33 (2015) 337–347. doi:10.1016/j.asoc.2015.04.046.

[36] F.E. Sandnes, L. Neyse, Y.-P. Huang, Simple and practical skin detection with static RGB-color lookup tables: A visualization-based study, in: IEEE Int. Conf. Syst. Man, Cybern., 2016: pp. 2370–2375.

[37] S. Jairath, S. Bharadwaj, M. Vatsa, R. Singh, Adaptive Skin Color Model to Improve Video Face Detection, in: Mach. Intell. Signal Process., Springer, 2016: pp. 131–142.

[38] W. Song, D. Wu, Y. Xi, Y.W. Park, K. Cho, Motion-based skin region of interest detection with a real-time connected component labeling algorithm, Multimed. Tools Appl. (2016) 1–16.

[39] A. Gupta, A. Chaudhary, Robust skin segmentation using color space switching, Pattern Anal. Image Anal. 26 (2016) 61–68.

[40] M.M. Oghaz, M.A. Maarof, A. Zainal, M.F. Rohani, S.H. Yaghoubyan, A hybrid Color space for skin detection using genetic algorithm heuristic search and principal component analysis technique, PLoS One. 10 (2015). doi:10.1371/journal.pone.0134828.

[41] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2017). doi:10.1109/TPAMI.2016.2644615.

[42] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2015. doi:10.1007/978-3-319-24574-4_28.

[43] C.Ó. Conaire, N.E. O'Connor, A.F. Smeaton, Detector adaptation by maximising agreement between independent data sources, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2007. doi:10.1109/CVPR.2007.383448.

[44] A. Cheddad, J. Condell, K. Curran, P. Mc Kevitt, A skin tone detection algorithm for an adaptive approach to steganography, Signal Processing. 89 (2009) 2465–2478. doi:10.1016/j.sigpro.2009.04.022.

[45] Y.H. Chen, K.T. Hu, S.J. Ruan, Statistical skin color detection method without color transformation for real-time surveillance systems, Eng. Appl. Artif. Intell. 25 (2012) 1331–1337. doi:10.1016/j.engappai.2012.02.019.

[46] M. Kawulok, Fast propagation-based skin regions segmentation in color images, in: 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition., 2013. doi:10.1109/FG.2013.6553733.

[47] M. Kawulok, J. Kawulok, J. Nalepa, Spatial-based skin detection using discriminative skin-presence features, Pattern Recognit. Lett. 41 (2014) 3–13. doi:10.1016/j.patrec.2013.08.028.

[48] N. Brancati, G. De Pietro, M. Frucci, L. Gallo, Human skin detection through correlation rules between the YCb and YCr subspaces based on dynamic color clustering, Comput. Vis. Image Underst. 155 (2017) 33–42. doi:10.1016/j.cviu.2016.12.001.

[49] Q. Zhu, C.-T. Wu, K. Cheng, Y. Wu, An adaptive skin model and its application to objectionable image filtering, in: Proc. 12th Annu. ACM Int. Conf. Multimed., 2004: p. 56. doi:10.1145/1027527.1027538.

[50] J. Ruiz-Del-Solar, R. Verschae, Skin detection using neighborhood information, in: Proc. - Sixth IEEE Int. Conf. Autom. Face Gesture Recognit., 2004: pp. 463–468. doi:10.1109/AFGR.2004.1301576.

[51] S.L. Phung, A. Bouzerdoum, D. Chai, Skin segmentation using color pixel classification: Analysis and comparison, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 148–154. doi:10.1109/TPAMI.2005.17.

[52] S.J. Schmugge, S. Jayaram, M.C. Shin, L. V. Tsap, Objective evaluation of approaches of skin detection using ROC analysis, Comput. Vis. Image Underst. 108 (2007) 41–51. doi:10.1016/j.cviu.2006.10.009.

[53] J. Stöttinger, A. Hanbury, C. Liensberger, R. Khan, Skin paths for contextual flagging adult videos, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2009: pp. 303–314. doi:10.1007/978-3-642-10520-3_28.

[54] L. Huang, T. Xia, Y. Zhang, S. Lin, Human skin detection in images by MSER analysis, 18th IEEE Int. Conf. Image Process. (2011) 1257–1260. doi:10.1109/ICIP.2011.6115661.

[55] W.R. Tan, C.S. Chan, P. Yogarajah, J. Condell, A Fusion Approach for Efficient Human Skin Detection, Ind. Informatics, IEEE Trans. 8 (2012) 138–147. doi:10.1109/TII.2011.2172451.

[56] J.C. Sanmiguel, S. Suja, Skin detection by dual maximization of

detectors agreement for video monitoring, Pattern Recognit. Lett. 34 (2013) 2102–2109. doi:10.1016/j.patrec.2013.07.016.

[57] J.P.B. Casati, D.R. Moraes, E.L.L. Rodrigues, SFA: A human skin image database based on FERET and AR facial images, in: IX Work. Visao Comput. Rio Janeiro, 2013.

[58] M.R. Mahmoodi, S.M. Sayedi, F. Karimi, Z. Fahimi, V. Rezai, Z. Mannani, SDD: A skin detection dataset for training and assessment of human skin classifiers, in: Knowledge-Based Eng. Innov. (KBEI), 2015 2nd Int. Conf., 2015: pp. 71–77.

[59] L. Nanni, A. Lumini, Skin Detection for Reducing False Positive in Face Detection, in: V.M. Petrova (Ed.), Adv. Eng. Res. Vol.16, Nov. Publ. 2017, Nova Publisher, 2017.

https://www.novapublishers.com/catalog/product_info.php?products_id=60326&osCsid=.

[60] A. Angelova, Y. Abu-Mostafa, P. Perona, Pruning training sets for learning of object categories, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2005: pp. 494–501. doi:10.1109/CVPR.2005.283.

[61] D.M.W. Powers, Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation, J. Mach. Learn. Technol. 2 (2011) 37–63. doi:10.1.1.214.9232.

[62] J.H. McDonald, Handbook of Biological Statistics (3rd ed.), Sparky House Publ. Balt. Maryl. (2014). doi:10.1017/CBO9781107415324.004.

SUPPLEMENTAL MATERIAL: Extended version of Table 3 (including results for different thresholds)

		Dataset									AP	Rank
Perf. Indicator		F_1 measure										
Method	τ / w_τ	Feeval	Prath.	MCG	UChile	Compaq	SFA	HGR	Schmugge	VMD	FnF	
GMM		0.513	0.581	0.688	0.615	0.600	0.789	0.658	0.595	0.130	91.590	12
SPL	-2.5	0.549	0.537	0.615	0.562	0.492	0.726	0.845	0.501	0.302	87.580	
	-2	0.544	0.551	0.621	0.568	0.494	0.700	0.845	0.490	0.321	89.380	
	-1.5	0.544	0.567	0.627	0.565	0.497	0.628	0.845	0.482	0.341	90.290	13
	-1	0.529	0.578	0.624	0.554	0.494	0.583	0.828	0.462	0.354	91.180	
	-0.5	0.512	0.582	0.615	0.543	0.485	0.521	0.801	0.435	0.367	92.140	
Bayes	50	0.552	0.672	0.702	0.605	0.613	0.473	0.711	0.508	0.264	95.710	
	70	0.564	0.663	0.711	0.643	0.617	0.593	0.793	0.547	0.272	94.680	
	90	0.567	0.649	0.706	0.660	0.609	0.686	0.840	0.565	0.268	93.100	7
	110	0.565	0.631	0.694	0.661	0.599	0.760	0.871	0.569	0.252	92.380	
	140	0.547	0.599	0.667	0.642	0.574	0.830	0.885	0.555	0.221	90.510	
Cheddad	75	0.561	0.633	0.668	0.563	0.615	0.281	0.357	0.571	0.209	93.710	
	100	0.619	0.641	0.704	0.637	0.608	0.468	0.596	0.601	0.274	92.860	
	125	0.596	0.597	0.667	0.649	0.588	0.683	0.767	0.571	0.261	89.200	10
	150	0.518	0.513	0.585	0.567	0.567	0.550	0.850	0.384	0.151	80.700	
	175	0.439	0.448	0.526	0.484	0.540	0.581	0.812	0.349	0.067	75.110	
Chen		0.287	0.540	0.656	0.598	0.549	0.791	0.732	0.571	0.165	87.200	14
SA1	225	0.486	0.579	0.639	0.511	0.536	0.809	0.752	0.453	0.087	76.980	
	200	0.535	0.603	0.657	0.553	0.580	0.823	0.787	0.479	0.153	83.860	
	175	0.558	0.613	0.664	0.567	0.593	0.788	0.768	0.482	0.199	86.770	11
	150	0.575	0.615	0.668	0.571	0.599	0.730	0.732	0.471	0.234	88.500	
	100	0.560	0.576	0.645	0.543	0.588	0.572	0.614	0.440	0.276	90.780	
SA2	30	0.524	0.683	0.739	0.656	0.641	0.758	0.842	0.580	0.169	91.660	
	40	0.519	0.690	0.749	0.661	0.644	0.760	0.831	0.590	0.163	91.770	
	50	0.515	0.693	0.755	0.663	0.645	0.771	0.806	0.594	0.156	91.910	8
	85	0.513	0.665	0.749	0.647	0.627	0.799	0.508	0.584	0.130	92.100	
	120	0.497	0.626	0.714	0.623	0.591	0.801	0.354	0.553	0.108	91.480	
SA3	125	0.506	0.638	0.715	0.575	0.591	0.897	0.498	0.511	0.101	86.640	
	100	0.521	0.674	0.738	0.592	0.613	0.896	0.667	0.535	0.114	87.930	
	75	0.536	0.699	0.755	0.608	0.633	0.886	0.785	0.566	0.129	89.150	6
	50	0.539	0.709	0.762	0.625	0.647	0.863	0.877	0.586	0.147	89.800	
	25	0.544	0.707	0.752	0.631	0.647	0.817	0.893	0.591	0.172	90.980	
DYC		0.588	0.599	0.680	0.663	0.618	0.569	0.616	0.613	0.275	88.710	9
SegNet		0.717	0.730	0.813	0.802	0.737	0.889	0.869	0.708	0.328	99.978	3
U-Net		0.576	0.787	0.779	0.713	0.686	0.848	0.836	0.671	0.332	98.848	4
Vote1	1.25	0.594	0.717	0.754	0.670	0.666	0.737	0.849	0.625	0.269	92.980	
	1.5	0.594	0.717	0.754	0.670	0.666	0.737	0.849	0.625	0.269	92.980	
	1.75	0.593	0.684	0.733	0.684	0.635	0.792	0.872	0.611	0.250	92.980	5
	2	0.593	0.684	0.733	0.684	0.635	0.792	0.872	0.611	0.250	92.980	
Vote2	1.25	0.689	0.809	0.836	0.826	0.737	0.717	0.921	0.697	0.432	99.780	
	1.5	0.728	0.82	0.838	0.83	0.768	0.820	0.946	0.697	0.476	99.780	
	1.75	0.736	0.811	0.816	0.81	0.772	0.854	0.949	0.700	0.481	99.780	2
	2	0.648	0.755	0.778	0.755	0.751	0.892	0.949	0.700	0.328	99.780	
Vote3	1.25	0.705	0.825	0.829	0.795	0.760	0.800	0.930	0.690	0.457	99.800	
	1.5	0.73	0.821	0.84	0.829	0.774	0.869	0.949	0.708	0.428	99.800	
	1.75	0.733	0.812	0.841	0.829	0.773	0.902	0.950	0.714	0.423	99.800	1
	2	0.647	0.755	0.817	0.806	0.750	0.914	0.949	0.706	0.330	99.800	