Skin detection by dual maximization of detectors agreement for video monitoring[☆]

Juan C. SanMiguel[a,*], Sergio Suja[a]

[a] *Video Processing and Understanding Lab, Escuela Politécnica Superior,*
Universidad Autónoma of Madrid, E-28049 Madrid, Spain

Abstract

This paper presents an approach for skin detection which is able to adapt its parameters to image data captured from video monitoring tasks with a medium field of view. It is composed of two detectors designed to get high and low probable skin pixels (respectively, regions and isolated pixels). Each one is based on thresholding two color channels, which are dynamically selected. Adaptation is based on the agreement maximization framework, whose aim is to find the configuration with the highest similarity between the channel results. Moreover, we improve such framework by learning how detector parameters are related and proposing an agreement function to consider expected skin properties. Finally, both detectors are combined by morphological reconstruction filtering to keep the skin regions whilst removing wrongly detected regions. The proposed approach is evaluated on heterogeneous human activity recognition datasets outperforming the most relevant state-of-the-art approaches.

Keywords: Skin detection, Detector adaptation, Color space selection, Performance optimization

1. Introduction

Detecting skin regions in images that correspond to human body parts is an important task in many areas such as human-computer interaction, gesture analysis and content-based image retrieval. Recently, recognition of human activities in video has become a relevant topic where the detection and tracking of body parts, via skin detection, plays a key role [1][2]. However, such detection faces many challenges related to the scenario (illumination changes and backgrounds with skin-like surfaces), the field of view (medium-small skin areas) and the limited availability of training data, which decrease the performance of traditional skin detection approaches for this recognition task. An adaptable detector to scenario conditions (requiring few training data) is therefore needed.

Traditional skin detection consists on two stages namely training and classification phases [3]. First, the input color space (RGB) of the images is transformed to get a better representation of the skin color distribution only using chrominance information (the luminance one is usually dropped). Then, second phase performs skin classification over such distribution through parametric (e.g., Neural Network [4]), non-parametric (e.g., Naive Bayes histograms [5]) and explicit methods (e.g., manual thresholding [6]). There exist many studies addressing combinations among color spaces and state-of-the-art classifiers [7][8][9][10]. It is commonly agreed that detection performance increases by transforming input data to cylindrical color spaces (HSV, HSI). However, different conclusions are found for the use of the luminance component as some studies indicate that it might improve detection [8] whereas others do not [10]. Among classifier methods, Random Forest obtained the best performance [11] due to its good generalization properties and low requirements of training data. Moreover, detectors can be combined to improve overall detection by sequentially removing false detections [12]. The main drawback of traditional approaches is the static modeling (using large training sets) that can not be efficiently adapted to each test condition.

Adaptation of skin detection has been usually targeted to address illumination changes. Skin illumination dependency can be removed through color constancy correction [7]. However, recent results [8] show that accuracy of well-known corrections, such as gray-world and skin locus, is data-dependent and, therefore, performance improvement can not always be guaranteed. Another adaptation consists on combining previously trained (or global) skin models with local models extracted from high probable skin pixels of each considered condition such as the adaptations of Hue [13] and Bayes [14] histogram-based models. Moreover, local models can be also computed via the color of detected face regions [15][16]. However, these approaches require large amounts of training data for global models and manual setting of the combination factor, where non-adequate values might introduce errors of local models into the final skin detection. The threshold-based Gaussian Mixture Model (GMM) [5] is extended in [17] where high-confidence regions for skin and non-skin

[☆] This work has been partially supported by the Spanish Government (TEC2011-25995 EventVideo).

[*] Corresponding author

Email addresses: juancarlos.sanmiguel@uam.es (Juan C. SanMiguel), sergio.suja@estudiante.uam.es (Sergio Suja)

are obtained via detectors (for faces and people). Then, these regions are used for threshold optimization and mixture weight adaptation. However, [17] has some limitations as it uses large training sets, heuristically defines the optimization function and assumes accurate detection, which can not be always guaranteed. Unlike previous work, detector adaptation can be achieved by maximizing its agreement with other independent detector through their output similarity or agreement [18]. Hence, both detectors will have high and low agreement for, respectively, correct and false skin detections. Demonstrated over explicit thresholding on the three HSV channels, it obtained very promising results without requiring global or local modeling [18]. However, there exist problems with skin-like background surfaces and low percentages of skin data in the image, which make the adapted thresholds tend to increase the number of false positives. Finally, adaptation could be also in the modeling aspect, where optimum color spaces are selected [19] or mixtures of their channels [20] to maximize skin detection performance. However, both approaches are limited as they require large amounts of training data for accurate modeling (as they employ histograms), use maximum a posteriori criteria over test data [19] or propose heuristics over train data that can not be generalized to all situations [20].

In this paper, we propose a skin detection approach that adapts its parameters and selects optimum color channels for image data captured from video monitoring tasks with a medium field of view, where few training data are available containing medium-small skin regions. It consists on two detectors that use channels of different color spaces for getting high and medium probable skin pixels (respectively, isolated pixels and compact regions). Each detector is based on a two-channel thresholding scheme which is adapted by maximizing the agreement (i.e., similarity) between the two-channel results. Relations between thresholds are learned through parametric kernels, which improve the maximization process. Finally, both detectors are combined through morphological opening by reconstruction that selects the best regions indicated by high probable skin pixels. The results demonstrate that the proposed approach outperforms the related state-of-the-art on public human activity recognition datasets.

The rest of the paper is organized as follows. Section 2 defines the maximization of agreement. Section 3 describes the proposed approach. Experiments are shown in section 4 and the conclusions are presented in section 5.

2. Theoretical framework

In the proposed approach, skin detectors are adapted to data within an optimization framework [18][21] that selects the best detectors's configuration based on agreement maximization (AM). It consists of three basic elements (detectors applied, agreement measure and optimization process) which are described as follows.

For each detector, data sources are employed (i.e., color channels) to detect skin pixels by explicitly defining the boundary of the cluster containing them as:

$$D_i(\mathbf{x}) = \left\{ \begin{array}{lll} 1 & if & \tau_i^{inf} < C_i(\mathbf{x}) < \tau_i^{sup} \\ 0 & & otherwise \end{array} \right. , \quad (1)$$

where C_i is the ith color channel (data source), $\{\tau_i^{inf}, \tau_i^{sup}\}$ are its thresholds and $\mathbf{x}=(m,n)$ are the 2D pixel coordinates. Typically, each detector uses two color channels ($i=\{1,2\}$) obtaining two results (D_1 and D_2), whose binary outputs are combined using logical AND.

Then, the aim is to online adjust the thresholds of each channel by maximizing the similarity between D_1 and D_2, which is measured using an agreement function $A(D_1, D_2)$ such as mutual information [18] or signal correlation [21]. This function should give high or low values when both results, respectively, agree or disagree at pixel level (i.e., they do not have the same binary output). For finding the optimum parameters that maximize the agreement, a standard optimization algorithm can be used such as simplex [18] or gradient ascent [21], which is defined as:

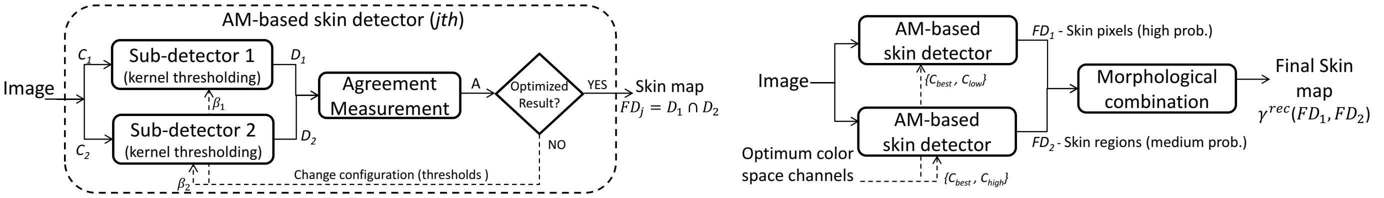
$$S_j = S_{j-1} + \eta \nabla \tilde{A}(S_{j-1}), \quad (2)$$

where $S_j = \{\tau_{j1}^{inf}, \tau_{j1}^{sup}, \tau_{j2}^{inf}, \tau_{j2}^{sup}\}$ is the parameter set that controls the skin detector at jth iteration, $\tilde{A}$ is $A(D_1, D_2)$ using S_j to obtain the results $\{D_1, D_2\}$, $\nabla\tilde{A}(S_j)$ is the gradient of $\tilde{A}$ particularized at S_j and $\eta > 0$ is a constant. $\nabla\tilde{A}(S_j)$ can be obtained by evaluating $\tilde{A}(S)$ in a neighborhood of S_j. This process is iteratively repeated until the parameter set does not change (reaching the optimum values with maximum agreement).

However, this framework has no constraints in the parameter optimization process which makes the thresholds tend to increase the number of false positives or negatives as agreement is high in certain non-desirable situations (e.g., both D_1 and D_2 binary masks with a high number of 1's or 0's), which is critical when dealing with high or low quantity of skin pixels in the image. This four-parameter optimization might be slow if many values are tested. Moreover, there is no indication of which channels of color spaces are better for increasing the agreement and complex combination schemes can be designed considering the properties of the employed detectors. The detector shall be able to automatically adapt to different conditions (skin proportions, skin-like surfaces, lighting) whilst selecting the optimum color space channels.

3. Dual maximization for skin detection

We propose an approach to detect skin in single images of human activity recognition videos where, for each image, it dynamically selects the best configuration starting from a predefined one. First, we introduce such adaptation using the AM framework [18][21], whose optimization is



(a) (b)

Fig. 1. Block diagrams of the proposed (a) detector based on agreement maximization (AM) and (b) framework for skin detection in images.

improved by learning parameter relations through kernel-thresholding and including a new agreement measure (Fig. 1(a)). Then, two AM-based detectors are composed to detect skin-like regions and high-probable skin pixels (via optimal selection of color space channels), which are later combined using binary morphology (Fig. 1(b)) for maximizing performance. They are described as follows.

3.1. AM-based skin detector

Kernel-based thresholding. First, we aim to constrain the values of the thresholds $\{\tau_i^{inf}, \tau_i^{sup}\}$ in the optimization process of the AM-based detectors. In certain situations, these thresholds tend to similar or max/min values (e.g., $\tau_i^{inf} = 0$ and $\tau_i^{sup} = 1$, if $C_i(\mathbf{x}) \in [0,1]$) as the agreement is maximized in both cases. We propose to solve this problem by learning their relation using training data and performing kernel-based thresholding, which is defined as:

$$D_i(\mathbf{x}) = \begin{cases} 1 & if \quad K(C_i(\mathbf{x})) > \beta_i \\ 0 & \quad otherwise \end{cases}, \qquad (3)$$

where D_i is the skin detection for the C_i color channel, $C_i(\mathbf{x})$ is the value at position $\mathbf{x}$, $K(\cdot)$ is a kernel function and $\beta_i \in [0,1]$ is a threshold that influences the presence of false positives in the skin detection. The objective of $K(\cdot)$ is to define the relation between the lower (τ_i^{inf}) and upper (τ_i^{sup}) thresholds as well as their bounds (maximum and minimum values), which avoids to use incorrect values during the optimization. In addition, $K(\cdot)$ allows to compute the probability of $C_i(\mathbf{x})$ being in a range $[\tau_i^{inf}, \tau_i^{sup}]$ depending on just one threshold (β_i) dynamically adapted to image data. As kernel function, we can use any parametric or non-parametric function that describes the distribution of skin pixels in the selected color space channel. For simplicity, we use a Gaussian kernel:

$$K(p) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(p-\mu)^2}{2\sigma^2}} \qquad (4)$$

where $\{\mu, \sigma\}$ are the mean and standard deviation of the skin distribution obtained with training data. Note that few training data is required as we do not aim to precisely model skin data with $K(\cdot)$ (only the approximate relation between τ_i^{inf} and τ_i^{sup}). Moreover, the optimization process is sped up as the number of parameters is reduced from four (Eq. 1) to two (Eq. 3) for each detector.

After obtaining the optimum parameter values, detector output FD_j is generated by a logical AND combination of its corresponding two channel analysis D_i as follows:

$$FD_j = D_1 \cap D_2 \qquad (5)$$

Detector agreement. For the agreement function, we first compute the expected skin pixel proportion (SP_k) for each scenario to analyze using training data:

$$SP_k = \frac{s_k}{M_k \cdot N_k}, \qquad (6)$$

where s_k is the number of annotated skin pixels for kth image and $\{M_k, N_k\}$ are its size. For each scenario, we use a small set of K training images to get the maximum and minimum values of SP_k (SP_{max} and SP_{min}).

Then, we propose a correlation-based agreement function to compute the similarity between the results $\{D_1, D_2\}$ when their respective skin proportions fall within the expected proportion range of the scenario being analyzed:

$$A(D_1, D_2) = \begin{cases} \rho(D_1, D_2) & if \quad max(SP_1, SP_2) < w \cdot SP_{max} \quad \cap \\ & \qquad min(SP_1, SP_2) > \frac{SP_{min}}{w} \\ 0 & \qquad otherwise \end{cases}, \qquad (7)$$

where $\{SP_1, SP_2\}$ are the skin proportions of the sub-detector $\{D_1, D_2\}$ using Eq. 6, $\{SP_{max}, SP_{min}\}$ are the maximum and minimum skin proportions obtained from training data, w is a weighting factor to consider higher or lower proportions than the training data ones and $\rho(\cdot,\cdot)$ is the Pearson product-moment correlation coefficient [22]:

$$\rho(D_1, D_2) = \left| \frac{E\left[(D_1 - \mu_{D_1})(D_2 - \mu_{D_2})\right]}{\sigma_{D_1}\sigma_{D_2}} \right|, \qquad (8)$$

where $\{\mu_{D_1}, \mu_{D_2}\}$ and $\{\sigma_{D_1}, \sigma_{D_2}\}$ are, respectively, the means and standard deviations of $\{D_1, D_2\}$; $E[\cdot]$ is the expectation; $\rho \in [0,1]$, with values close to 1 indicating high correlation (agreement) between both outputs.

This new agreement function $A(D_1, D_2)$ improves existing ones [18][21] by enforcing a similar number of detected skin pixels in both results while penalizing false detections and proportions out of the expected range.

3.2. Optimum channel selection

For each type of scenario, we obtain the best color space channels among the most popular ones (RGB, HSV, YCbCr and Lab) to detect skin pixels by determining their discriminative capabilities over the training data.

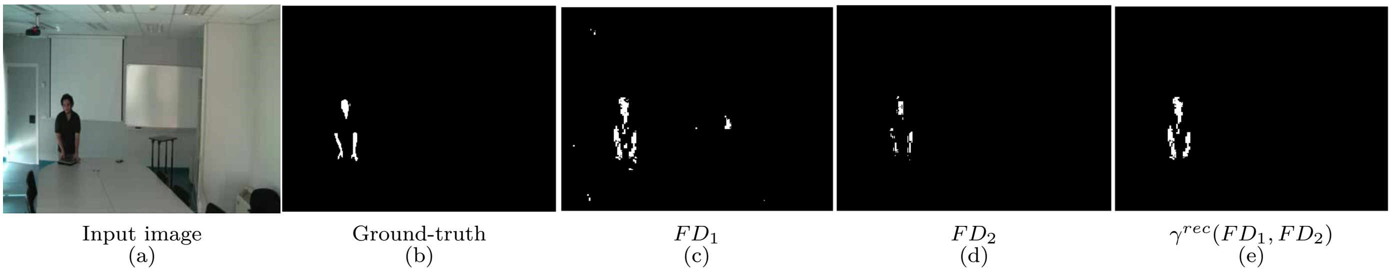


Fig. 2. Sample results for an image of the EDds dataset showing the output of the skin detectors FD_1(H-a, $\beta_H = .06$, $\beta_a = .16$) and FD_2 (H-b, $\beta_H = .10$, $\beta_b = .49$) after optimum channel selection and their combination through mathematical morphology γ^{rec}.

As histogram modeling has proven to be an accurate method for analyzing color spaces [7][19][20], we use it to get a simple skin detector for each channel by a normalized histogram computed using all the available training images. Then, we apply these (channel-based) simple detectors to the same training images by taking an acceptance probability of 0.01 (getting most of the skin pixels in the train images) and compute the following mean ratio:

$$r_i = \frac{1}{K}\sum_{k=1}^{K}\frac{d_{ik}}{s_k}, \tag{9}$$

where d_{ik} and s_k are the number of, respectively, detections (for the ith channel) and annotated skin pixels in kth training image; and K is the number of training images.

Then, we have experimentally observed a similar true positive rate for all channel detectors (i.e., recall) as they have the same acceptance probability and, therefore, we select the channel whose detector has lowest ratio r_i (i.e., minimum false positive rate) as the optimum one:

$$C_{best} = \underset{i}{argmin}\left(r_i\right). \tag{10}$$

As at least two channels are required in the AM framework, we propose to set C_{best} as the base one and choose, among the rest of the channels, those that satisfy the precision objectives of each detector. Then, we design two AM-based detectors to get skin pixels with high precision (i.e., isolated pixels) or low precision (i.e., compact regions). For high precision detection, we use the Pearson correlation ρ (Eq. 8) to select the channel that gives minimum agreement with the optimum channel (i.e., few pixels are detected, with high probability):

$$C_{high} = \underset{i}{argmin}\left(\rho(H_{C_{best}}, H_i)\right), \tag{11}$$

where $\{H_{C_{best}}, H_i\}$ are the binary outputs of the histogram-based skin detector using, respectively, the color channels C_{best} and C_i. For low precision detection (and high recall of skin pixels), we similarly select the channel that gives maximum agreement with the optimum channel (i.e., several low probable pixels are detected obtaining regions, where non-skin regions might appear in the detection):

$$C_{low} = \underset{i}{argmax}\left(\rho(H_{C_{best}}, H_i)\right). \tag{12}$$

Finally, the two detectors are conformed by using the color channels $\{C_{best}, C_{low}\}$ for FD_1 and $\{C_{best}, C_{high}\}$ for FD_2. Fig. 2 shows an example of such selection and the obtained results after the agreement maximization process. The detector FD_1 uses the channels H-a which is tuned to get skin-like regions (see Fig. 2(c)) whereas the other detector FD_2 combines H-b to obtain high probable skin pixels that belong to skin regions (see Fig. 2(d)).

3.3. Detector combination

After selecting the optimum channels of the AM-based skin detectors and optimizing their parameters, they are combined to improve the final result of the skin detection. Considering that one of the detectors obtains highly probable pixels whereas the other gets compact skin-like regions (that might correspond to skin or similar objects), we propose to use a morphological reconstruction filter [23] to retain only the skin regions marked by the highly probable pixels of FD_2(among all the regions of FD_1).

Morphological filters by reconstruction allow to eliminate undesirable details in binary images without affecting the structure of the desirable ones. As we aim to preserve the skin regions indicated by FD_2, we apply an opening by reconstruction of erosion, $\gamma^{rec}(X;Y)$, which requires two elements (input image X and marker Y) to get the result (reconstructed image). As input image, we use the output of FD_1 (containing skin and non-skin regions) whereas the output of FD_2 for the marker (high probable skin pixels pointing the skin regions). For applying such filtering, we first define the conditional dilation operation as:

$$Y_p = (Y_{p-1} \oplus se) \wedge X, \tag{13}$$

where X is the output FD_1, $\oplus$ is the standard morphological dilation, se is a square structuring element whose size depends on the image size and Y_p is the reconstructed image at pth iteration. For $p = 1$, we use the erosion of the marker image FD_2 with se as the initial result ($Y_1 = FD_2 \Theta se$). Note that this conditional dilation is computed recursively. Then, the reconstructed image is defined as:

$$\gamma^{rec}(X;Y) = Y_\infty, \tag{14}$$

which requires to define stability conditions for stopping this infinite iterative analysis. We set the maximum number of iterations (1000) as well as the minimum difference among consecutive iterations ($||Y_p - Y_{p-1}|| < 1e^{-6}$) as such conditions. Finally, a conditional closing morphological filter is used for performing hole filling over the reconstructed image. Standard closing operation consists on a dilation followed by an erosion and we extend it by conditioning the addition of new pixels (filled holes) to their existence in the results of the detectors (i.e., they are marked as skin pixels in FD_1 or FD_2). This operation allows to add pixels (of the holes) to the skin output avoiding the inclusion of wrong pixels (i.e., non increasing the false positive rate). The output of the proposed skin detector is the reconstructed image filtered by this modified closing. Fig. 2 shows an example of such combination where the selected skin regions are kept in the final results (Fig. 2(e)).

4. Experiments

We present the results of the proposed[1] and related approaches for detecting skin pixels on human activity recognition datasets where such detection is an important task.

4.1. Setup

As evaluation set, we have selected images from public datasets for human activity recognition: EDds[2], LIRIS[3], SSG[4], UT[5], and AMI[6]. This set covers a wide variety of situations, viewing distances and resolutions (ranging from 320x240 to 720x576) where skin detection has many challenges due to, among others, illumination changes or poor visibility. For each dataset, around 50 images have been selected and the corresponding ground truth has been manually generated at pixel level. In total, 290 images compose the evaluation set containing more than 870000 skin pixels, which have been equally divided into two sets for training (~450000) and testing (~420000). Note that we are not using large-scale datasets [5] as the focus is on limited training data and the analysis of human activity recognition datasets. Both conditions are satisfied for the video monitoring domain.

For comparison purposes, we have selected the explicit thresholding with fixed values for the H-S (T_HS) and Cb-Cr (T_CbCr) channels as defined in [6], the Bayesian-based method which thresholds the ratio of belonging to the skin and non-skin distributions using histograms (BAY_H) [5] and Gaussian Mixture Models (BAY_G) [5], and the Random Forest method applied over HSV [11]. For BAY_H and BAY_G, the threshold is empirically chosen to maximize performance. As adaptive approaches, we use the global-local adaptation of the H channel (ASD) [13] and the maximization of mutual information using a co-training scheme (MMI) over H-S [18]. All approaches are implemented in MATLAB and use the default parameters indicated by their respective authors.

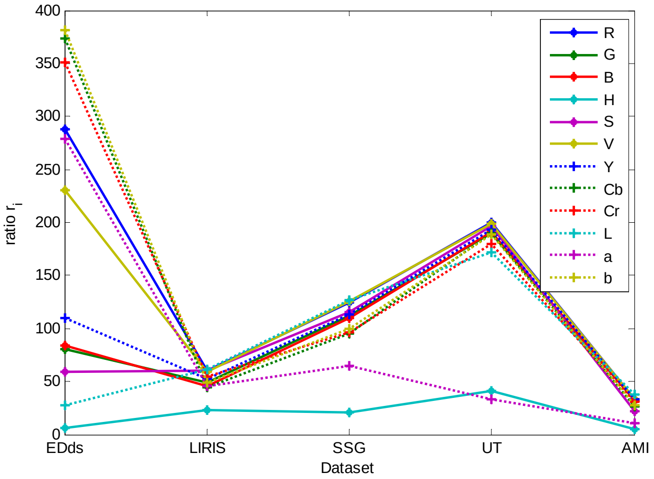


Fig. 3. Results for base optimum channel selection over the selected datasets using mean detection ratio (Eq. 9).

For evaluating detection performance, we use standard Precision (P), Recall (R) and F-score (F):

$$P = TP/(TP + FP), \tag{15}$$

$$R = TP/(TP + FN), \tag{16}$$

$$F = 2 \cdot P \cdot R/(P + R). \tag{17}$$

where TP, FP and FN are, respectively, the correct, wrong and missed skin detections. Finally, we set the following parameters of the proposed approach: $w = 4$ (deviation of the trained skin proportion, see Eq. 7) and the size of the structuring element se proportional to the image size (3×3 for 320×240 image sizes, see Eq. 13).

4.2. Optimum channel selection

Fig. 3 and Table 1 summarize the results for optimum channel selection. Fig. 3 depicts the mean detection ratio (Eq. 9) of the histogram-models computed for each channel over the training set. As it can be observed, the minimum value corresponded to the H channel (of HSV) for all the datasets except for UT, where *a* channel (of Lab) obtained best results, closely followed by H. This indicates that H channel had stable results and high discriminative power (for skin and non-skin regions) in the considered scenarios. These results agree with previous works [7][8][10], where cylindrical color spaces (HSV, HSI) provided the best performance with other types of images.

[1] Detailed description of the evaluation set, annotated data, additional results and software implementations can be found at http://www-vpu.eps.uam.es/publications/SkinDetDM

[2] http://www-vpu.eps.uam.es/DS/EDds/

[3] http://liris.cnrs.fr/harl2012/

[4] http://www-vpu.eps.uam.es/publications/SkinDetDM

[5] http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

[6] http://corpus.amiproject.org/

	EDds	LIRIS	SSG	UT	AMI
R	.081	.070	**.040**	**.024**	.109
G	.108	.227	.042	.046	.129
B	.116	.261	.049	.035	.127
H	-	-	-	**.606**	-
S	.094	**.057**	.135	.065	**.070**
V	.063	.103	.430	.031	.106
Y	.087	.178	.055	035	.121
Cb	.046	.241	.148	.140	.228
Cr	.049	.159	.146	.101	.157
L	.096	.101	.120	.098	.105
a	**.126**	**.337**	**.334**	-	**.298**
b	**.040**	.230	.121	.163	.146

Table 1: Pearson correlation (Eq. 8) between each color space channel and the base optimum one selected for each datasets. Maximum and minimum correlation values are bold marked.

Table 1 shows the Pearson correlation (Eq. 8) with the optimum channel determined for each dataset. For the two detectors that compose the proposed approach, channels with maximum correlation were selected for recognizing high probable pixels (detector FD_2) whereas minimum correlation values indicate that the result contains skin-like regions (detector FD_1). We have obtained the following combinations for EDds (H-a and H-b), LIRIS (H-a and H-S), SSG (H-a and H-R), UT (a-H and a-R) and AMI (H-a and H-S) datasets.

Fig. 4 illustrates visual examples of the detectors. The first case considers an image from EDds with the following configurations for FD_1 ($\beta_H = .06$, $\beta_a = .16$, $A_{H-a} = .82$) and FD_2 ($\beta_H = .10$, $\beta_b = .49$, $A_{H-b} = .25$). In the second case, the image is from the LIRIS dataset so detector configuration changes for FD_1 ($\beta_H = .26$, $\beta_a = .15$, $A_{H-a} = .41$) and FD_2 ($\beta_H = .58$, $\beta_S = .90$, $A_{H-S} = .03$). Third example uses SSG dataset obtaining a different detector configuration for FD_1 ($\beta_H = .35$, $\beta_a = .15$, $A_{H-a} = .81$) and FD_2 ($\beta_H = .50$, $\beta_R = .49$, $A_{H-R} = .05$). Last case is from AMI dataset with the configuration for FD_1 ($\beta_a = .25$, $\beta_H = .05$, $A_{a-H} = .23$) and FD_2 ($\beta_H = .95$, $\beta_S = .31$, $A_{H-S} = .10$). The channels and thresholds are dynamically adapted to each considered data. A common trend is observed in the four presented cases, channel combinations with higher agreements (detector FD_1) produce skin regions whereas lower agreements (detector FD_2) mainly provide an output with high probable skin pixels indicating which regions correspond to skin.

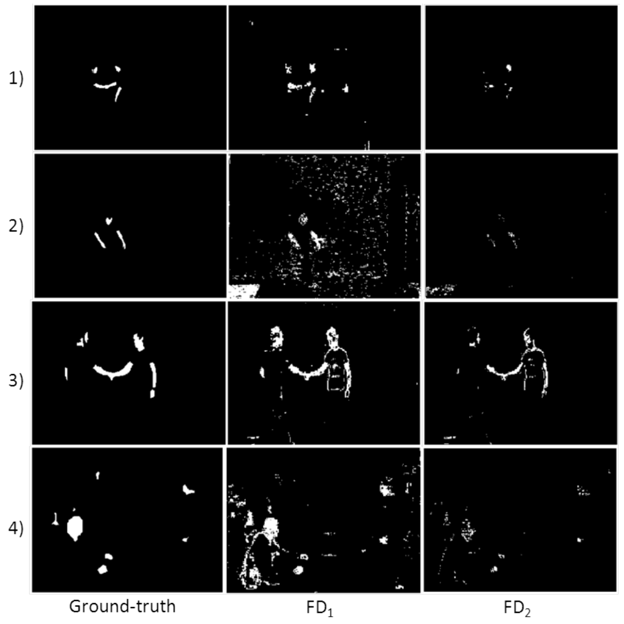


Fig. 4. Sample results of the detectors (FD_1 and FD_2) obtained after optimum channel selection.

4.3. Skin detection

Skin detection results of the proposed and compared approaches are presented in Table 2. As it can be observed, the results of the LIRIS dataset exhibited a clear decrease in performance for all the approaches as compared to the results of the other datasets. This is due the office furniture used in the dataset, which contains several skin-like surfaces affecting the precision results. It demonstrates that skin detection in such scenarios is complex and partially solved. In general, fixed-thresholding approaches (T_HS and T_CbCr) got medium performance showing that, albeit effective, the use of parameters with fixed values does not generalize well for a variety of heterogeneous scenarios. Differences between H-S and Cb-Cr are due to the better skin clustering properties of H-S for the selected data. BAY obtained good performance demonstrating that non-skin data can be efficiently used to improve final skin detection. In particular, the use of non-skin modeling allows to have high precision values as the approach is robust against skin-like surfaces. However, it depends on a decision threshold which its difficult tuning is critical to get the best performance (for the results, the optimum value is selected after testing with a set of thresholds ranging from 0 to 4). Between the two BAY versions, BAY_G is better than BAY_H as parametric modeling is more appropriate when dealing with small training sets. The recent RF approach similar performance among compared state-of-the-art obtaining a good precision-recall trade-off. As this machine learning approach also uses non-skin samples for skin modeling, it also benefits from having high precision values whereas keeping most of the skin (recall) as BAY_G and BAY_H. However, the three approaches (BAY_H, BAY_G and RF) suffer no-adaptability to each scenario and the particularities of each image (e.g., changing illumination) are treated as outliers of their skin models, thus limiting their performance. Adaptive approaches (ASD and MMI) presented very low performance indicating that introducing adaptive capabilities into skin detection is not an easy task. In some situations both approaches got high recall detecting most of the skin pixels but their precision was very low having high false detection rate as they produced detections covering several non-skin parts of the im-

Table 2: Comparison for selected skin detection approaches. Best results are bold marked. Last row indicates the Percentage increase (%) of each measure with respect to the best state-of-the-art performance.

Approach	EDds			LIRIS			SSG			UT			AMI			Mean		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
T_CbCr [6]	.253	.706	.373	.067	.914	.125	.148	.854	.252	.258	.839	.395	.242	.694	.359	.194	.801	.312
T_HS [6]	.398	.484	.437	.122	.327	.178	.385	.548	.453	.326	.571	.415	.396	.321	.354	.326	.450	.378
BAY_H [5]	.626	.502	.557	.147	.647	.239	**.515**	.493	.504	.330	.590	.423	.531	.804	.639	.430	.607	.503
BAY_G [5]	**.647**	.524	**.579**	.158	.690	.258	.469	.476	.472	.394	.455	.422	**.610**	.784	.686	.455	.586	.513
RF [11]	.502	.685	.580	.104	**.886**	.187	.436	.766	.558	.284	**.897**	.432	.503	**.930**	.653	.366	**.833**	.508
ASD [13]	.022	**.733**	.043	.038	.770	.072	.164	**.902**	.278	.002	.251	.004	.044	.531	.082	.054	.637	.100
MMI [18]	.055	.436	.099	.040	.800	.077	.056	.552	.101	.041	.141	.063	.020	.549	.039	.042	.496	.078
Proposed	.623	.648	**.636**	**.189**	.698	**.298**	.457	.754	**.569**	**.413**	.755	**.534**	.598	.842	**.699**	**.456**	.739	**.564**
%△ best	-0.03	-10.9	+9.0	+19.6	-21.2	+15.5	-15.2	-16.0	+1.9	+4.8	-15.8	+23.6	-2.1	-9.4	+1.8	+0.2	-11.2	+10.0

age. Although the proposed approach did not obtain the best precision or recall in most of the cases, it improved all the compared approaches using the F-score, which considers the balance between the precision and recall. Globally, an increase around 10% is observed over the best state-of-the-art approach BAY_G (last column of Table 2). It evidences that by adapting simple threshold-based detectors, the performance can be clearly improved. Fig. 5 illustrates detection examples for each dataset showing the previously discussed behavior and the preference of approach with highest F-score instead of only high precision or recall values.

Regarding mean computation time (in milliseconds per pixel), simple approaches are faster such as T_CbCr, T_HS and BAY_H (with respectively 0.0022, 0.0020 and 0.0044). Learning-based approaches present intermediate cost (0.0314 for BAY_G and 0.0450 for RF). Adaptive approaches have high costs (0.152 for ASD and 0.212 for MMI). A non-optimized implementation of the proposed approach employs 0.521 ms/pixels being the heaviest among compared approaches which benefit from optimized code. The main reason regards the computation of the agreement function during the iterative optimization (Eq. 7). However, this cost can be dramatically reduced by applying integral images [18] or other optimization approaches.

5. Conclusions

In this paper, we have presented an adaptive approach for skin detection in images. It proposes to combine two detectors where one is designed to get skin-like regions whereas the other is tailored to get high probable skin pixels. Both detectors are combined through mathematical morphology to effectively keep the skin regions. Moreover, they are based on the agreement maximization framework, which computes the similarity between the sub-detectors of each detector and maximizes it in order to get better performance. This framework is extended to model the relation between the sub-detector parameters, to consider agreement within the expected ranges and to select the optimal channels of color spaces. Experimental results demonstrate the adaptive capabilities of the proposed approach improves performance of parameter-fixed, adaptive and learning-based state-of-the-art approaches.

As future work, we will explore the adaptive estimation of skin proportions for the agreement function and the application of the proposed approach to video sequences exploiting temporal relations between the frames.

References

[1] M. Ryoo, J. Aggarwal, Semantic representation and recognition of continued and recursive human activities, Int. journal of computer vision 82 (1) (2009) 1–24.

[2] J. C. SanMiguel, J. M. Martínez, A semantic-based probabilistic approach for real-time video event recognition, Computer Vision and Image Understanding 116 (9) (2012) 937–952.

[3] A. Elgammal, C. M., D. Hu, Skin detection-a short tutorial, Encyclopedia of Biometrics (2009) 1–10.

[4] C. Lin, Face detection in complicated backgrounds and different illumination conditions by using ycbcr color space and neural network, Pattern Recognition Letters 28 (16) (2007) 2190–2200.

[5] M. Jones, J. Rehg, Statistical color models with application to skin detection, Int. Journal of Computer Vision 46 (1) (2002) 81–96.

[6] Y. Wang, B. Yuan, A novel approach for human face detection from color images under complex background, Pattern Recognition 34 (10) (2001) 1983–1992.

[7] P. Kakumanu, S. Makrogiannis, N. Bourbakis, A survey of skin-color modeling and detection methods, Pattern recognition 40 (3) (2007) 1106–1122.

[8] R. Khan, A. Hanbury, J. Stöttinger, A. Bais, Color based skin classification, Pattern Recognition Letters 33 (2) (2012) 157–163.

[9] S. Schmugge, S. Jayaram, M. Shin, L. Tsap, Objective evaluation of approaches of skin detection using roc analysis, Computer Vision and Image Understanding 108 (1) (2007) 41–51.

[10] J. Chaves-González, M. Vega-Rodríguez, J. Gómez-Pulido, J. Sánchez-Pérez, Detecting skin in face recognition systems: A colour spaces study, Digital Signal Processing 20 (3) (2010) 806–823.

[11] R. Khan, A. Hanbury, J. Stoettinger, Skin detection: A random forest approach, in: IEEE Int. Conf. on Image Processing (ICIP), 2010, pp. 4613–4616.

[12] B. Jedynak, H. Zheng, M. Daoudi, Skin detection using pairwise models, Image and Vision Computing 23 (13) (2005) 1122–1130.

[13] F. Dadgostar, A. Sarrafzadeh, An adaptive real-time skin detector based on hue thresholding: A comparison on two motion tracking methods, Pattern Recognition Letters 27 (12) (2006) 1342–1352.

[14] H. Sun, Skin detection for single images using dynamic skin color modeling, Pattern recognition 43 (4) (2010) 1413–1420.

[15] P. Yogarajah, J. Condell, K. Curran, A. Cheddad, P. McKevitt, A dynamic threshold approach for skin segmentation in color images, in: IEEE Int. Conference on Image Processing (ICIP), 2010, pp. 2225–2228.

[16] W. Tan, C. Chan, P. Yogarajah, J. Condell, A fusion approach for efficient human skin detection, IEEE Trans. on Industrial Informatics 8 (1) (2012) 138–147.

[17] T. Chen, P. Tan, L. Ma, M. Cheng, A. Shamir, S. Hu, Poseshop: human image database construction and personalized content synthesis, IEEE Trans on Visualization and Computer Graphics 19 (5) (2013) 824–837.

[18] C. Conaire, N. O'Connor, A. Smeaton, Detector adaptation by maximising agreement between independent data sources, in: IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–6.

[19] A. Albiol, L. Torres, E. Delp, Optimum color spaces for skin detection, in: IEEE Int. Conf. on Image Processing (ICIP), Vol. 1, 2001, pp. 122–124.

[20] G. Gomez, On selecting colour components for skin detection, in: Int. Conf. on Pattern Recognition, Vol. 2, IEEE, 2002, pp. 961–964.

[21] J. C. SanMiguel, J. M. Martínez, Shadow detection in video surveillance by maximizing agreement between independent detectors, in: IEEE Int. Conf. on Image Processing (ICIP), 2009, pp. 1141–1144.

[22] P. Chen, P. Popovich, Correlation: Parametric and non-parametric measures, Thousand Oaks, 2002.

[23] F. Shih, Image processing and mathematical morphology: Fundamentals and applications, CRC, 2009.

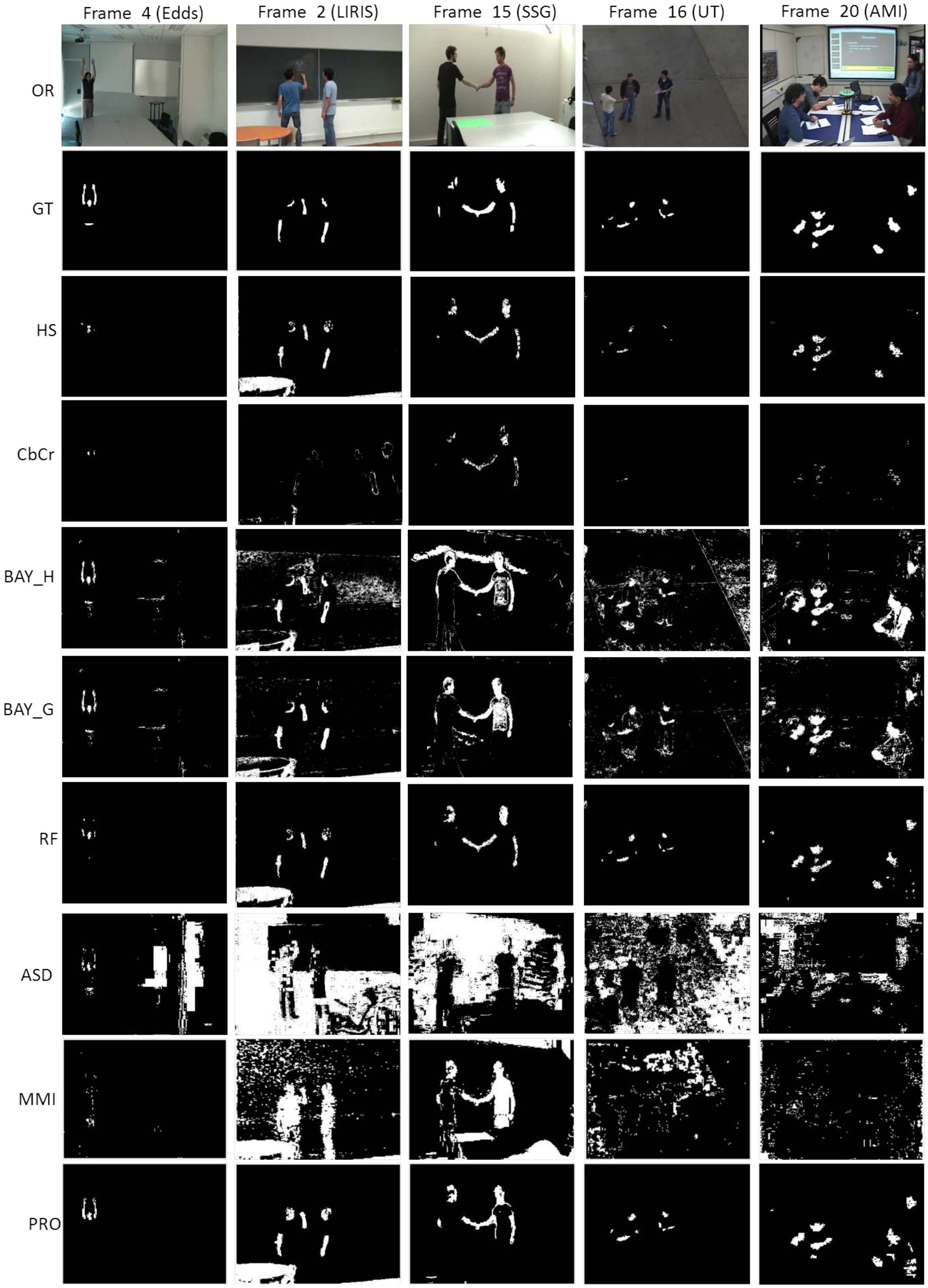


Fig. 5. Sample results of the compared skin detectors. Key. OR:Original image. GT:Ground-truth. T_HS:Threshold over HS [6]. T_CbCr:Threshold over CbCr [6]. BAY_H:Histogram-based Bayesian detector [5]. BAY_G:GMM-based Bayesian detector [5]. RF:Random Forest approach [11]. ASD:Adaptive skin detector [13]. MMI:Maximization of mutual information [18]. PRO:Proposed approach.