Skin segmentation using multiple thresholding

Francesca Gasparini, Raimondo Schettini*
DISCO (Dipartimento di Informatica, Sistemistica e Comunicazione)
Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy

ABSTRACT

The segmentation of skin regions in color images is a preliminary step in several applications. Many different methods for discriminating between skin and non-skin pixels are available in the literature. The simplest, and often applied, methods build what is called an "explicit skin cluster" classifier which expressly defines the boundaries of the skin cluster in certain color spaces. These binary methods are very popular as they are easy to implement and do not require a training phase. The main difficulty in achieving high skin recognition rates, and producing the smallest possible number of false positive pixels, is that of defining accurate cluster boundaries through simple, often heuristically chosen, decision rules. In this study we apply a genetic algorithm to determine the boundaries of the skin clusters in multiple color spaces. To quantify the performance of these skin detection methods, we use recall and precision scores. A good classifier should provide both high recall and high precision, but generally, as recall increases, precision decreases. Consequently, we adopt a weighted mean of precision and recall as the fitness function of the genetic algorithm. Keeping in mind that different applications may have sharply different requirements, the weighting coefficients can be chosen to favor either high recall or high precision, or to satisfy a reasonable tradeoff between the two, depending on application demands. To train the genetic algorithm (GA) and test the performance of the classifiers applying the GA suggested boundaries, we use the large and heterogeneous Compaq skin database.

Keywords: Skin segmentation, explicit skin cluster classifier, genetic algorithm.

1. INTRODUCTION

The segmentation of skin regions in color images is a preliminary step in several applications, such as image and video classification and retrieval in multimedia databases, semantic filtering of web contents (through the definition of medium-level features), human motion detection, human computer interaction, and video-surveillance. It is also useful in image processing algorithms, as well as in intelligent scanners, digital cameras, photocopiers, and printers. Many different methods for discriminating between skin and non-skin pixels are available in the literature. These can be grouped in three types of skin modeling: parametric, nonparametric, and explicit skin cluster definition methods. The Gaussian parametric models[1] assume that skin color distribution can be modeled by an elliptical Gaussian joint probability density function. Nonparametric methods estimate skin color distribution from the histogram of the training data without deriving an explicit model of skin color[2]. The simplest, and often applied, methods build what is called an "explicit skin cluster" classifier which expressly defines the boundaries of the skin cluster in certain color spaces. The underlying hypothesis of methods based on explicit skin clustering is that skin pixels exhibit similar color coordinates in an appropriately chosen color space. These binary methods are very popular as they are easy to implement and do not require a training phase. The main difficulty in achieving high skin recognition rates, with the smallest possible number of false positive pixels, is that of defining accurate cluster boundaries through simple, often heuristically chosen, decision rules. In this study we compare the performance of various explicit skin cluster methods applying the thresholds presented in the literature with that achieved when a genetic algorithm is applied to determine the boundaries of the skin clusters in multiple color spaces. To quantify the performance of these skin detection methods, we use recall and precision scores. Classification results are assigned as true positive (TP), false positive (FP) and false negative (FN). Recall is defined as the ratio between the number of skin pixels correctly classified and the total number of actual skin pixels (TP/(TP+FN)), while precision is defined as the ratio between the number of skin pixels correctly classified and the total number of pixels labeled as skin pixels by the skin detection method considered (TP/(TP+FP)).

* E-mail: gasparini@disco.unimib.it, schettini@disco.unimib.it

A good classifier should provide high recall and high precision, but generally, as recall increases, precision decreases. Consequently, we have adopted a weighted harmonic mean of precision and recall as the fitness function of the genetic algorithm. Keeping in mind that different applications may have sharply different requirements, the weighting coefficients can be chosen to offer either high recall or high precision, or to satisfy a reasonable tradeoff between the two depending on application demands.

To train the genetic algorithm and test the performance of the algorithms with the GA suggested boundaries, we have used the large and heterogeneous Compaq skin database[3]. Comparison of these results with those obtained applying the methods with the original boundaries, as reported in the literature, shows a significant improvement for all the algorithms considered.

1a

1b: YCbCr

1c: RGB

1d: HSV1

1e: HSV2

1f: HSI

1g: rgb

Figure 1. Examples of skin maps obtained applying the six methods considered. 1a: original image; 1b: YCbCr; 1c: RGB; 1d: HSV1; 1e: HSV2; 1f: HSI; and 1g: rgb. Note that some methods (YCbCr, and HSI) seem to be more recall oriented, others (HSV1 and rgb) more precision oriented, while still a third group (RGB and HSV) shows a good tradeoff between precision and recall.

2. SKIN SEGMENTATION

2.1 Binary skin classifiers

The methods considered in this paper separate skin and non skin colors using a piecewise linear decision boundary. These explicit skin cluster methods propose a set of fixed skin thresholds in a given color space. Some color spaces permit searching skin color pixels in the 2D chromatic space, reducing dependence on lighting variation, others, such as the RGB space, address the lighting problem by introducing different rules depending on illumination conditions (uniform daylight, or flash). Working within different color spaces, we have implemented the six different algorithms analyzed in this paper. They are named for the color space adopted: YCbCr[4], RGB[5], HSV1[6], HSV2[7], HSI[8] and rgb[9]. The details of their implementation can be found in the referenced papers and are summarized in the subsections here below. Examples of the skin maps obtained applying these methods to the image of Figure 1a are shown in Figures 1b-1g. Some of these methods, such as YCbCr (1b) and HSI (1f), are more recall oriented, some, such as HSV1 (1d) and rgb (1g), more precision oriented, while still others, such as RGB (1c) and HSV2 (1e), show a good tradeoff between recall and precision.

2.1.1 YCbCr

Chai and Ngan[4] develop an algorithm that exploits the spatial distribution characteristics of human skin color. A skin color map is derived and used on the chrominance components of the input image to detect pixels that appear to be skin. The algorithm then employs a set of regularization processes to reinforce those regions of skin-color pixels that are more likely to belong to the facial regions. We use only their color segmentation step here. Working in the YCbCr space the authors find that the ranges of Cb and Cr most representative for the skin-color reference map were:

$$77 \le Cb \le 127 \textbf{ and } 133 \le Cr \le 173 \textbf{.}$$

2.1.2 RGB

Kovac et al.[5] work within the RGB colour space and deal with the illumination conditions under which the image is captured. Therefore, they classify skin colour by heuristic rules that take into account two different conditions: uniform daylight and flash or lateral illumination.

Uniform daylight illumination:

$$R > 95, \quad G > 40, \quad B > 20$$
$$Max\{R,G,B\} - \min\{R,G,B\} < 15$$
$$|R - G| > 15, \quad R > G, \quad R > B$$

Flashlight or daylight lateral illumination:

$$R > 220, \quad G > 210, \quad B > 170$$
$$|R - G| \le 15, \quad B < R, \quad B < G\ .$$

2.1.3 HSV1

Tsekeridou and Pitas[6] work within the HSV color space and select pixels having skin-like colors by setting the following thresholds:

$$V \ge 40\ ;$$
$$0.2 < S < 0.6\ ;$$
$$0° < H < 25° \text{ or } 335° < H < 360°\ .$$

The selected range of H restricts segmentation ton reddish colors and the saturation range selected ensures the exclusion of pure red and very dark red colors, both of which are caused by small variations in lighting conditions. The threshold on V is introduced to discard dark colors.

2.1.4 HSV2

Starting from a training data set composed of skin color samples, Garcia and Tiziritas[7] compute the color histogram in hue-saturation-value (HSV) color space, and estimate the shape of this skin color subspace. They find a set of planes by successive adjustments depending on segmentation results, recording the equations shown below which define the six bounding planes found in the HSV color space case, where $H \in [-180° \; 180°]$:

$$V \geq 40$$
$$H \leq (-0.4\,V + 75)$$
$$10 \leq S \leq (-H - 0.1\,V + 110)$$
$$\text{if } H \geq 0 \quad S \leq (0.08\,(100 - V)\,H + 0.5\,V)$$
$$\text{if } H < 0 \quad S \leq (0.5\,H + 35)\,.$$

2.1.5 HSI

Hsieh et al.[8] use the HSI colour space system to design their colour classification algorithm because it is stable for skin colour under different lighting conditions. These rules apply to the intensity I, hue H and saturation S, and are detailed as follows:

$$I > 40$$
$$\textbf{if } 13 < S < 110\,, \quad 0° < H < 28° \text{ and } 332° < H < 360°$$
$$\textbf{if } 13 < S < 75\,, \quad 309° < H < 331°$$

The thresholds are empirically determined from the training set and the colour system transformation from RGB to HSI is defined as follows:

$$I_1 = \frac{1}{3}(R + G + B); \quad I_2 = \frac{1}{2}(R - B); \quad I_3 = \frac{1}{4}(2G - R - B).$$
$$I = I_1$$
$$S = \sqrt{I_2^2 + I_3^2}$$
$$H = \tan^{-1}\left(\frac{I_3}{I_2}\right)$$

2.1.6 rgb

Gomez and Morales[9] use a constructive induction approach to determine the skin map. Starting with the three rgb components in a normalized form and a simple set of arithmetic operators, the authors produce a model for skin detection. The algorithm uses a Restricted Covering Algorithm (RCA) as its selective learner. The RCA searches for single rules in parallel. Among the different combination rules presented by the authors, we have chosen the one with the highest precision and success rate:

$$\frac{r}{g} > 1.185\,, \quad \frac{r \cdot b}{(r + g + b)^2} > 0.107 \text{ and } \frac{r \cdot g}{(r + g + b)^2} > 0.112$$

where *rgb* are the normalized coordinates obtained as:

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B} \qquad b = \frac{B}{R+G+B}.$$

2.2 Genetic Algorithm

Genetic Algorithms (GAs) are adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetics. The GAs treat optimization problems as the competition of populations of evolving individuals (chromosomes), each considered as a candidate solution. A 'fitness' function evaluates each solution to decide whether it will contribute to the next generation. Then, through operations analogous to gene transfer in biological reproduction (selection of parents, crossover, replacement, and mutation), the algorithm produces a new population of candidate solutions.
Due to their random nature, GAs not only provide good solutions for optimization problems but also consistently outperform more traditional methods, improving the chances of finding a global solution. Whereas most stochastic search methods operate on the concept of a single solution, genetic algorithms operate on a population of solutions.

The main steps of GAs can summarized as follows:

1. Define a fitness function on which to base the criteria for the selection of those individuals of a population who will generate the next generation.
2. Define evolution strategies (selection of parents, crossover, replacement, mutation, and migration)
3. Randomly generate an initial population of solutions (chromosomes).
4. Compute and store the fitness for each individual in the current population
5. Generate the next population, selecting and evolving individuals from the previous population applying criteria based on survival of the fittest.
6. Repeat step 4 until a satisfactory solution is obtained.

We have applied a genetic algorithm to determine the boundaries of each of the six skin classifiers described here. The chromosomes that constitute each population are vectors of skin cluster boundary thresholds. Because a good classifier should have high recall and high precision, we adopted the following weighted harmonic mean of precision and recall, as the fitness function of the genetic algorithm:

$$fitness = \frac{recall \cdot precision}{\alpha \cdot recall + \beta \cdot precision} \tag{1}$$

The weighting coefficients α and β can be chosen to provide either high recall or high precision, or a reasonable tradeoff between the two depending on application demands.
The main GA settings used were:

- 30 chromosomes for each population
- Selection of parents: tournament with tournament size equal to 7, selecting the parents in proportion to their value in fitness.
- Crossover: inserting at random a separator in the homologous genes of the selected parents
- Mutation: with a probability of 0.01.

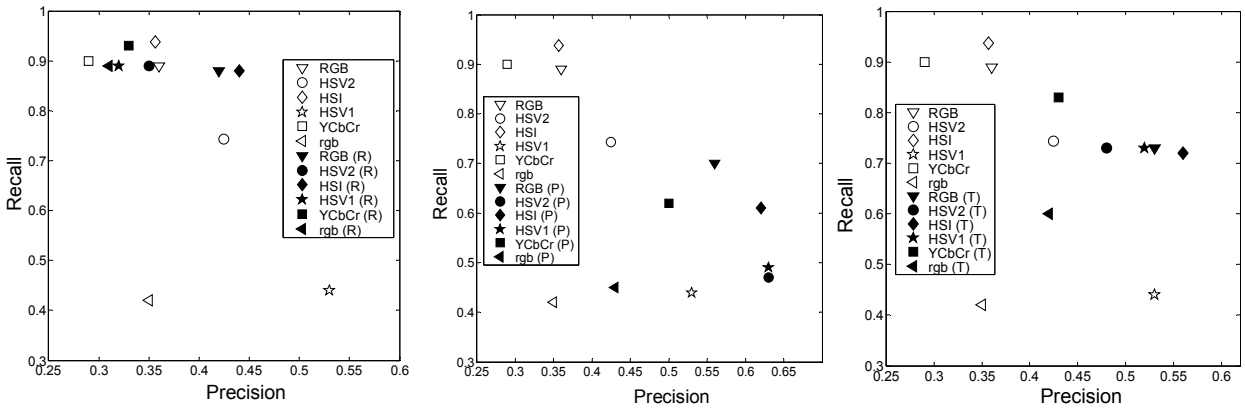
3. EXPERIMENTAL RESULTS AND DISCUSSION

We have evaluated the boundaries of the skin clusters by running the GA on a training set of 2.000.000 pixels, 295.147 of which skin, and the remaining 1.704.853 non-skin. These pixels were randomly chosen from some 2400 images taken from the Compaq skin database[3], which contains 24.043.259 skin pixels and 168.549.425 non-skin pixels.
For each classifier we performed three calculations: one to decide the thresholds in order to achieve high recall, setting α=0.2 and β=0.8 in the fitness function of equation 1, a second for high precision (α=0.8 and β=0.2) and a third for a reasonable tradeoff between these two measures (α=β=0.5).

The performance of the skin classifiers with the GA boundaries was then evaluated on all 2400 images of the database, computing true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). All these quantities refer to absolute values: TP, for example, represents the total number of skin pixels identified by the skin detector. From these measures, the recall and precision were evaluated for the three sets of boundary thresholds of each classifier.
In table 1 the values of recall and precision obtained with the genetic algorithm and the three different strategies described, are summarized and compared with those obtained using the thresholds recorded in the literature.
To further increase recall or precision, the coefficients of equation 1 can be exasperated. For example, setting α=0.1 and β=0.9 in the case of the HSI classifier, we obtained a recall of 93% and a precision of 39%.

Method	***Thresholds definition strategy***		***Recall (%)***	***Precision (%)***
YCbCr	Literature Thresholds		90	29
	GA Thresholds	Recall Oriented	93	33
		Precision Oriented	62	50
		Tradeoff	83	43
RGB	Literature Thresholds		89	36
	GA Thresholds	Recall Oriented	88	42
		Precision Oriented	70	56
		Tradeoff	73	53
HSV1	Literature Thresholds		44	53
	GA Thresholds	Recall Oriented	89	32
		Precision Oriented	49	63
		Tradeoff	73	52
HSV2	Literature Thresholds		74	42
	GA Thresholds	Recall Oriented	89	35
		Precision Oriented	47	63
		Tradeoff	73	48
HSI	Literature Thresholds		92	35
	GA Thresholds	Recall Oriented	88	44
		Precision Oriented	61	62
		Tradeoff	72	56
rgb	Literature Thresholds		42	35
	GA Thresholds	Recall Oriented	89	31
		Precision Oriented	41	67
		Tradeoff	60	42

Table 1. Recall and precision scores for the six classifiers, with respect to the values recorded in the literature and those obtained by the GA applying the different thresholding strategies: recall oriented (α=0.2 and β=0.8), precision oriented (α=0.8 and β=0.2) and tradeoff (α=β=0.5).



2a 2b 2c

Figure 2. Results of the experiments in terms of recall versus precision. Comparison of the six methods using the thresholds given in the literature (blank symbols) with the same methods using GA thresholds, (black symbols), applying in Figure 2a a recall oriented strategy, in Figure 2b a precision oriented strategy, and in Figure 2c a tradeoff strategy.

The results of these experiments are also visualized in Figure 2, in terms of recall versus precision. The results for the six methods applied using thresholds given in the literature (blank symbols) are compared with those using GA thresholds, (black symbols): applying in Figure 2a a recall oriented strategy, in Figure 2b a precision oriented strategy, and in Figure 2c a tradeoff strategy.

From the analysis of Table 1 and Figure 2, we can argue that:

- the performance of the methods studied indicates that different methods may be either more precision- or more recall-oriented. The qualification of "best method" is therefore application-dependent.

- the performance of the six methods considered, always improves with the GA boundaries, and significantly so for those (such as rgb and HSV1) with lower values in the literature:

    - In the case of recall oriented strategy (Figure 2a),all the methods show a strong increase in recall values, except for the HSI classifier, which already achieves a very high value of recall with the boundaries given in the literature, as well as the RGB classifier, which maintains the same recall value with the GA boundaries but shows a significant increase in the precision score.
      The GA boundaries tend to shift all the classifiers into the same region of the recall-precision plane (upper left hand corner), except for HSI and RGB methods.

    - In the case of precision oriented strategy (Figure 2b), all the methods increase in precision, moving towards the bottom right hand corner of the precision-recall plane. Those with high recall show a significant drop in that score, but still within an acceptable range, while those originally with low recall increase in both precision and recall.

    - In the case of the tradeoff strategy (Figure 2c), all the methods reach a good tradeoff between recall and precision, moving towards the upper right hand corner of the precision-recall plane. The worst classifier in this sense seems to be the rgb, registering the lowest value of recall with the lowest precision, but showing nevertheless substantial improvement with respect to values recorded in the literature.

- the HSI, as proposed by the literature and shown in Table 1, remains the best method in terms of recall.

As a final remark, we note that a good balance of the weighting coefficients of the fitness function permits the fine tuning of the methods with respect to recall or precision, to meet application demands.

REFERENCES

1. M.-H Yang and N. Ahuja, “Gaussian Mixture Model for Human Skin Colour and its Applications in Image and Video Databases”, SPIE/EI&T Storage and Retrieval for Image and Video Databases (San Jose, January 1999), pp. 458-466.
2. M. Jones, J. Rehg, “Statistical Color Models with Application to Skin Detection”, IEEE Conference on Computer Vision and Pattern Recognition, CVPR '99 (1999), pp. 274-280.
3. Compaq Cambridge Research Lab image-database. M. Jones and J. Rehg, “Statistical Colour Models with Application to Skin Colour Detection”. Compaq Cambridge Research Lab Technical Report CRL 98/11 (1998).
4. D. Chai and K. N. Ngan, Face segmentation using skin colour map in videophone applications, IEEE Transactions on Circuits and Systems for Video Technology 9 (4) (1999) 551-564.
5. J. Kovac, P. Peer and F. Solina, 2D versus 3D colour space face detection, 4th EURASIP Conference on Video/Image Processing and Multimedia Communications, Croatia, 2003, pp. 449-454.
6. S. Tsekeridou and I. Pitas, “Facial feature extraction in frontal views using biometric analogies", Proc. of the IX European Signal Processing Conference, vol. I, 315-318 (1998).
7. C.Garcia and G. Tziritas, Face detection using quantized skin colour regions merging and wavelet packet analysis, IEEE Transaction on Multimedia 1 (1999) 264-277.
8. I-S. Hsieh, K-C. Fan, and C. Lin, A statistic approach to the detection of human faces in colour nature scene, Pattern Recognition 35 (2002) 1583-1596.
9. G. Gomez and E. F. Morales, “Automatic feature construction and a simple rule induction algorithm for skin detection, Proc. Of the ICML workshop on Machine Learning in Computer Vision, A. Sowmya, T. Zrimec (eds), 31-38 (2002).