Adaptive skin segmentation via feature-based face detection

DOI:
10.1117/12.2052003

Adaptive skin segmentation via feature-based face detection

Michael J. Taylor, Tim Morris
School of Computer Science, University of Manchester, UK

ABSTRACT

Variations in illumination can have significant effects on the apparent colour of skin, which can be damaging to the efficacy of any colour-based segmentation approach. We attempt to overcome this issue by presenting a new adaptive approach, capable of generating skin colour models at run-time. Our approach adopts a Viola-Jones feature-based face detector, in a moderate-recall, high-precision configuration, to sample faces within an image, with an emphasis on avoiding potentially detrimental false positives. From these samples, we extract a set of pixels that are likely to be from skin regions, filter them according to their relative luma values in an attempt to eliminate typical non-skin facial features (eyes, mouths, nostrils, etc.), and hence establish a set of pixels that we can be confident represent skin. Using this representative set, we train a unimodal Gaussian function to model the skin colour in the given image in the normalised rg colour space – a combination of modelling approach and colour space that benefits us in a number of ways. A generated function can subsequently be applied to every pixel in the given image, and, hence, the probability that any given pixel represents skin can be determined. Segmentation of the skin, therefore, can be as simple as applying a binary threshold to the calculated probabilities. In this paper, we touch upon a number of existing approaches, describe the methods behind our new system, present the results of its application to arbitrary images of people with detectable faces, which we have found to be extremely encouraging, and investigate its potential to be used as part of real-time systems.

Keywords: face detection, colour distribution modelling, skin segmentation, adaptive skin segmentation

1. INTRODUCTION

As computer systems and electronic devices become increasingly ubiquitous, and increasingly capable, the development of more sophisticated and useful human-computer interfaces becomes not only more necessary, but also more viable. In the realm of computer vision, the capacity for a machine to detect a person, or people, in a still image, or a video stream, and be able to subsequently interact with them is of paramount importance to a number of different potential applications. Among others, these applications include face detection for person recognition or expression estimation, hand detection for gesture recognition, and large-scale person counting for crowd size estimation and surveillance.

Skin segmentation from within images is an extremely popular mechanism towards interfacing with people, largely because of skin's prominence where human beings are concerned, and because its appearance often conforms to certain rules, especially in terms of its colour. However, the colour distribution of skin, as it appears in an image, is the combination of its chromaticity and its illumination, and, as such, will vary greatly from person to person, and from circumstance to circumstance (in terms of environment, capturing device characteristics, and other such external factors).

Traditionally, colour-based skin detection has been achieved through the definition of decision rules, which specify acceptable values for the appearance of skin within a certain colour space. These decision rules are often determined empirically, whereby researchers will use large datasets to find optimal bounds. Although simple to interpret and quick to apply, these models are too rigid to be applied as global skin detection solutions. Various learning techniques have also been applied to the problem, which successfully eliminate the biases inherent to static decision rule models, but they are still highly dependent on their training data being representative of potential inputs to be successful. It is for these reasons that adaptive methods have risen to prominence, as the capacity for a segmentation system to self-calibrate, and tailor its classification rules specifically for its inputs, leading to much greater accuracies, is of extreme value.

We have achieved adaptation through the adoption of a feature-based face detector, as our segmentation process begins with the detection of faces within the given image. We then define sub-regions within the detected face regions to extract pixels from, in order to eliminate background information. As some of these pixels may still represent non-skin features, we filter them according to their relative luma values in an attempt to leave only skin pixels within the set. We then build a two-dimensional Gaussian function within the normalised rg colour space with our filtered pixel set, which we can then

use to calculate the probabilities of certain colours representing skin within the rest of the given image. Using a threshold in conjunction with these probabilities will result in the successful segmentation of skin.

In the remainder of this paper, we will look at a wide range of previous works in the field and discuss their various strengths and weaknesses (Section 2), describe the design and development process of our new approach (Section 3), report results pertaining to the quality of our system's segmentations and its capacity to be used in real-time (Section 4), and, finally, draw a number of conclusions from our findings and discuss potential future extensions to our work (Section 5).

2. PREVIOUS WORK

There are four classes of previous work that are relevant to our interests in this instance: static skin colour distribution models; non-parametric colour modelling; parametric colour modelling; and face detection-based skin colour modelling techniques.

2.1 Static skin colour distribution models

Skin colour distribution models that do not adapt to given inputs are referred to as "static", and generally define bounds within a colour space, which, when applied to a pixel in a given image, will return a binary classification on whether that pixel represents skin. Any pixel that falls within the given model's bounds will be positively classified, and vice-versa. There are a number of different static skin colour models, specified within a number of different colour spaces.

2.1.1 RGB

Due primarily to the simplicity with which it can be interpreted, RGB is a popular starting point for any colour-based segmentation system, and skin-related systems are no different [1,2,3,4]. RGB represents colours as combinations of red, green, and blue amounts of light, typically in the form of three 8bit channels. The most widely applied RGB skin model was developed by Kovac et al. [1], and it is defined as follows:

$$pixel = \begin{cases} skin, & R > 95 \; and \; G > 40 \; and \; B > 20 \; and \\ & \max\{R,G,B\} - \min\{R,G,B\} > 15 \; and \\ & |R - G| > 15 \; and \; R > G \; and \; R > B \\ non - skin, & otherwise \end{cases} \quad (1)$$

This decision rule can be quickly applied across all pixels in a given image, allowing for rapid skin classification, making it an attractive option for real-time systems. Its overall accuracy, however, is diminished by its mixing of chrominance and luminance data, meaning that even minor changes in illumination will result in changes to skin colour values. Furthermore, there is high correlation between the channels [2], meaning that the space represents large amounts of redundant information.

2.1.2 Normalised rg

Normalised rg is a normalisation of the RGB colour space, where the two components represent the proportion of red and green in the given colour, rather than the absolute intensity of them. Given that the process of normalisation dictates that the normalised values r, g, and b sum to 1, the third component (b) can simply be dropped to allow for a two-dimensional space, without any loss of information. A number of works have touched upon its application to skin detection [5,6], and Soriano et al. [7] developed the following skin colour equation:

$$pixel = \begin{cases} skin, & g < (-1.8423r^2 + 1.5294r + 0.0422) \; and \\ & g > (-0.7279r^2 + 0.6066r + 0.1766) \; and \\ & (r - 0.33)^2 + (g - 0.33)^2 > 0.004 \\ non - skin, & otherwise \end{cases} \quad (2)$$

Normalised rg is a strong candidate for optimal colour space for any systems concerned with real-time processing, as it allows for relatively computationally efficient classifications, as well as granting the natural efficiency benefits of working within a two-dimensional space rather than a three-dimensional space. Additionally, it has been shown that the skin colours of people of different ethnicities tend to be represented very similarly by normalised rg [8], meaning that models developed within the space can be very widely applied. However, although the conversion to colour proportions does negate the effects of illumination variation to a large degree, the space does not handle colours approaching black particularly well, as small colour changes will translate to large fractional changes.

2.1.3 HSV

An alternative colour space for skin segmentation represents colours in terms of their hues, saturations, and values. This is an interesting colour space, as, unlike RGB, it functions similarly to human colour perception [2]. Ideally for the purposes of skin segmentation, the separation of the 'value' parameter (representing "brightness") affords some degree of resistance to illumination variation, and also allows for focus to be placed exclusively on the chrominance parameters (hue and saturation). Extensive experimentation within the colour space carried out by Sobottka and Pitas [9] lead to the definition of the following skin cluster:

$$pixel = \begin{cases} skin, & 0 \leq H \leq 50 \; and \; 0.23 \leq S \leq 0.68 \\ non-skin, & otherwise \end{cases} \quad (3)$$

Although studies have shown that HSV-based skin colour models are capable of producing impressive results [5,10,11], the space often presents a number of complications that can limit its appeal. Unlike normalised rg, different skin colours will tend to be represented by relatively different clusters, meaning that the development of globally applicable skin colour models is largely infeasible. This concept was highlighted by the model described in Eq. 3, where the authors themselves stated that, while the model was adequate for the segmentation of Asian and Caucasian skin, its capacity to segment African skin was unproven [9].

2.1.4 YCbCr

Used widely in digital video compression systems, YCbCr is an encoding of the RGB colour space. Colours are represented by a combination of their "luma" (computed via a weighted sum, approximating the response of human eyes to colour brightness) and blue and red chromatic components. Similarly to HSV, the separation of the brightness component is of great benefit where skin colour modelling is concerned. A number of previous works have explored the space's potential to be used for skin segmentation [5,12,13,14], and the experimental results of Hu et al. [15] demonstrate that the decision rule specified by Eq. 4 is capable of effectively extracting skin of all types.

$$pixel = \begin{cases} skin, & 137 < Cr < 177 \; and \\ & 77 < Cb < 127 \; and \\ & 190 < Cb + 0.6Cr < 215 \\ non-skin, & otherwise \end{cases} \quad (4)$$

CbCr space models, such as this one, have been successfully applied to skin segmentation tasks [13,15], highlighting their legitimacy to be used for any skin colour-related system. It has also been proven that, similarly to the normalised rg space, skin colours derived from people of various ethnicities are very tightly clustered within the CbCr space [16], meaning that individual models can successfully be applied to people of any skin tone. Despite the separation of the luminance component, skin colours can have greatly varying CbCr representations given different illuminations, very similarly to our findings in the normalised rg space (see Section 3), which puts a limitation on the usefulness of static clusters within the space, such as the one detailed by Eq. 4.

Static skin colour models are popular because of the simplicity of their implementations, and the speed at which they can allow a system to operate, making them particularly attractive for real-time requirements. Additionally, they can be used without the requirement of a training process. However, the empirical derivation of static skin colour models introduces certain degrees of subjectivity, and can result in the "overfitting" of models to experimental data. Furthermore, the relationships between the conditions of each model are complex enough that making adjustments for specific needs is practically infeasible.

2.2 Non-parametric skin colour distribution modelling

Non-parametric colour modelling methods estimate the probabilities that certain colours represent skin from given skin and non-skin training data. Such models usually exist in the form of look-up tables, attributing skin probabilities to specific colours, without the derivation and definition of an explicit colour cluster. Histogram-based approaches are common implementations of this form of classifier [2,5,7], where the given colour space will be quantised into an appropriate number of bins, each of which represents a particular combination of discrete colour components. These bins will store the number of times individual colour combinations occur within the given training data, and be subsequently normalised to give a probability distribution. Such a distribution can also be achieved through the application of the Bayes' rule described by Eq. 5, wherein the prior probabilities *P(skin)* and *P(¬skin)* can be estimated from the overall numbers of samples in the training sets [5,17].

$$P(skin|c) = \frac{P(c|skin)P(skin)}{P(c|skin)P(skin)+P(c|\neg skin)P(\neg skin)} \quad (5)$$

In general, non-parametric methods are simple to train and apply to images, and look-up tables negate the necessity of a model to "fit" the colour distribution shape of any input. However, depending upon the complexity of the colour space being worked within, the tables can often require significant storage space [2]. Additionally, unlike their parametric counterparts, non-parametric methods are incapable of generalising or interpolating data or colours, meaning that the samples used for training must be perfectly representative of any potential input for successful segmentation to occur.

2.3 Parametric skin colour distribution modelling

Parametric colour modelling methods represent colour distributions such that they can be described by a small number of parameters, which will typically be derived from the colour distribution of a set of positive skin samples. Parametric methods offer, in most cases, extremely compact colour distribution representations, which can be massively beneficial if storage space and computational complexity are a concern, making them of great interest to a number of works [6,13,17]. A skin colour distribution can be modelled by a Gaussian joint-probability density function, defined as:

$$p(c|skin) = \frac{1}{2\pi|\Sigma_s|^{1/2}} \cdot e^{-\frac{1}{2}(c-\mu_s)^T \Sigma_s^{-1}(c-\mu_s)} \quad (6)$$

Here, c is a colour vector (a particular combination of colour component values), and μ_s and Σ_s are the Gaussian distribution parameters (representing the mean colour vector and the covariance matrix, respectively) for colour model s, which are calculated from the given training data. Therefore, Eq. 6 can be used to calculate the probability of input colour vector c representing skin. Another major benefit parametric methods afford is in their capacity to generalise training data, and, as such, they can interpolate colours which may not have necessarily been represented explicitly by the samples. However, validation of parametric models is a concern, as how well the model "fits" is dependent upon the shape of the colour distribution within the given colour space. Some researchers have argued that single Gaussian functions are insufficient for accurate distribution modelling, and have instead opted for Gaussian mixture models [18] or "elliptical boundary models" [19]. It has been shown by Terrillon and Akamatsu [20], on the other hand, that the normalised colour spaces (rg in particular) yield very good fits using merely unimodal Gaussian models, due primarily to the elliptical colour clusters that skin tones exhibit within those spaces.

2.4 Face detection-based skin colour distribution modelling

In an effort to develop adaptive methods, capable of generating models that have been calibrated to given inputs, a number of researchers have explored the concept of using feature-based face detectors to sample skin regions. Such sampling will allow for estimations of skin colour distributions within images to be made, allowing for specifically tailored classifiers. Most works in this field have adopted the detector developed by Viola and Jones [21], which can be attributed to the combination of its high accuracy and relatively low computational complexity. This detector uses a novel cascade structure of "weak" classifiers, which are designed to detect Haar-like features within images. Such a structure means that negative samples can be discarded quickly and efficiently, whereas samples that are more likely to represent faces can be processed more thoroughly, so that positive classifications can be made with strong confidence. Additionally, the detector makes use of an "integral image" system, which allows for rapid image sub-region calculations, and affords great computational efficiency.

Mittal et al. [22] proposed a hand detection approach that involved the combination of the results of a global, generalised skin detector, with those derived from the Viola-Jones face detector [21]. The approach then uses spatial information in order to complete the segmentation. Although they claim segmentation accuracies greater than accepted state-of-the-art methods, the use of spatial techniques renders the approach rather unsuitable for real-time applications, as the processing of a single 640x360-pixel image would take in the region of two minutes. Wimmer and Radig [23] developed a system that utilises the same face detector [21], but uses detection region of interest (ROI) data to extract a small set of pixels with a skin mask. A skin mask is derived from a threshold applied to an empirically determined probability map, which has been trained to estimate the likelihood of specific coordinates of a detected face ROI representing skin. Extracted pixels are then used to build a parametric colour model to be applied to the entire image. While interesting, we do not believe that this approach's pixel extraction process is ideal, and it may be prone to undersampling in some cases, and over-sampling in others (especially when a subject has hair that may be occluding their forehead), both of which are capable of harming the efficacy of any generated skin colour model.

Again using the Viola-Jones detector [21], Hsieh et al. [24] developed a system that would allow for accurate hand detection, by generating skin colour models from detected face data. Given a square detection region, an inner face region is defined as being a square centered on the same point, but being only 0.6 of the original's size in both width and

height (granting a sample of 36% of the pixels of the detection). The pixels within this region are then filtered according to their luminance, whereby the symmetric property of Gaussian distributions is applied to remove "dark" pixels from the set. The remaining pixels are then used to build decision rules in the hybrid normalised rg/R (r,g,R) space, which will simply classify a given pixel as skin if it is within two standard deviations of the mean value for all three components. Given what we know of skin colour's elliptical clustering in the normalised rg space [20], however, we do not believe that orthogonal, binary decision boundaries are the ideal choice for accurate segmentation.

A new face detection algorithm, based upon directional Sobel edges, was developed by Liao and Chi [25]. Within detected face regions, they sample pixels from a small, predefined window, the location of which has presumably been empirically derived, in an effort to consistently extract "good" skin pixels from the right-hand cheeks of subjects. A histogram of the extracted pixels within the hue domain is then computed, and non-zero local minima, both greater and smaller than the peak hue, are found. These minima stand as the upper and lower bounds of skin segmentation, respectively, as every pixel in the given image is classified according to their hue value. The window used by this approach samples only 4% of the given detected face, which we believe could cause critical undersampling in a high proportion of cases. Additionally, we do not believe that hue alone is sufficient for accurate skin colour modelling in anything but the most ideal illumination conditions, a belief reinforced by the sheer number of works that have felt it necessary to combine hue information with saturation information in order to model skin colour [5,9,10,11].

Despite the stated shortcomings of these approaches, they all report results better than those returned by select non-adaptive methods. However, the increases in segmentation accuracy granted by adaptive methods will almost always come at the price of computational complexity, as the process of applying a model to an image must be preceded by the process of actually building or calibrating the model itself. Careful integration of system components is critical to maintaining efficiency, and ensuring that benefits to accuracy come at no less than a reasonable computational cost.

3. OUR APPROACH

The development of our approach stems from observations made on the colour distributions (in the normalised rg space) of the skin of people within lecture theatres. We were interested in discovering the degree to which distributions could vary, given large numbers of people (of numerous ethnicities) within environments having different lighting conditions (in terms of the numbers of light sources, their colours, their intensities, etc.). Our findings were profound, as we discovered that there could potentially be absolutely no overlap between clusters derived from images that could be reasonably labelled as representing ideal lighting conditions, meaning that the phenomenon can be the result of even non-extreme circumstances. A sample of the findings that illustrate this point can be seen in Fig. 1.

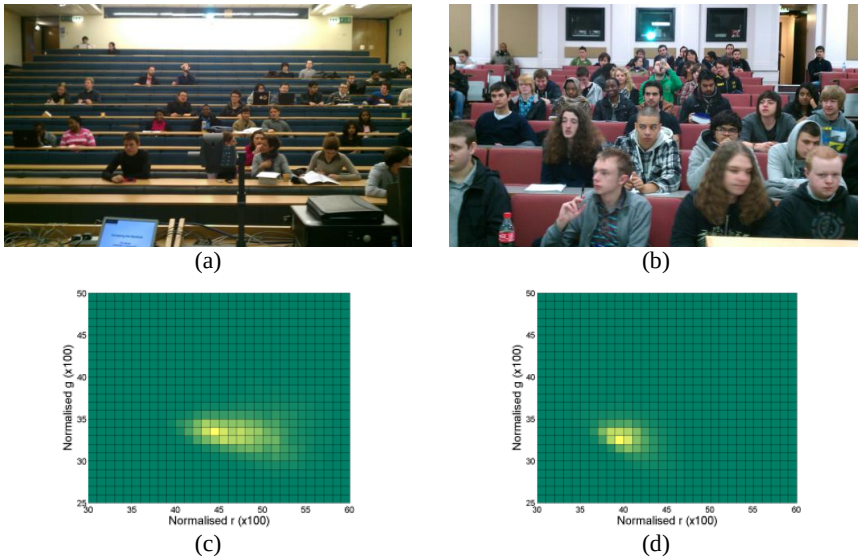


Figure 1: (a)(b) Our original lecture theatre scenes, and (c)(d) the respective colour distributions (in the normalised rg space) of the skin manually segmented from within them.

From such findings, we can draw a number of conclusions. Firstly, given what we know about the mathematics of the normalised rg colour space, it is likely that we would have found similar results in most other spaces (a small amount of

experimentation in the hue-saturation space, for example, demonstrated somewhat similar cluster variation). Secondly, despite the radical changes in cluster size, orientation, and location within the normalised rg space observed, the distributions all retain a generally elliptical shape. Thirdly, if such a scenario is possible given large numbers of people (granting wide colour distributions) in reasonably well-illuminated environments, then it is even more likely given individuals in unconstrained environments (the typical use-case for any segmentation technique). Finally, we can determine that any pre-trained classifier would perform poorly for images such as these, as, potentially, given the absence of any distribution overlap, any colour correctly classified as skin in one of the above images, for example, will be incorrectly positively classified in the other, and vice-versa. Furthermore, if, for instance, the manually segmented skin of Fig. 1(b) was used as the training data for a colour model, which was then applied to Fig. 1(a), the detection accuracy would be virtually nil.

These conclusions, combined with our perceived benefits and shortcomings of the aforementioned adaptive approaches (see Section 2.4), have had great influences on the development of our system, the process pipeline of which can be seen in Fig. 2.

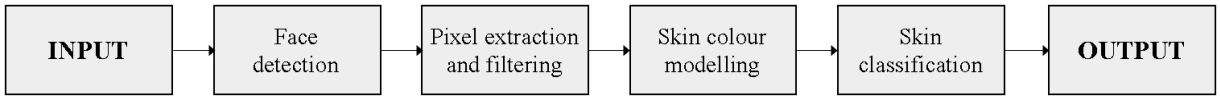


Figure 2: A simple overview of our system.

3.1 Feature-based face detection

In order for any process to adapt to its input, a method of sampling must exist. In our case, we require a sample of a person's (or people's) skin colour. For a colour-adaptive system, it stands to reason that we would want a colour-invariant sampling method, and feature-based techniques would, therefore, be ideal. Given that arbitrary pictures of people are more likely to display faces than any other detectable body part, and that the face is the most feature-rich part of the body (despite that being a double-edged sword when selecting candidate "skin" pixels, as some previous work has demonstrated [21,25]), reliable sampling would be best served through face detection.

As far as feature-based face detection is concerned, the Viola-Jones detector represents a good choice for adoption. As well as its innovations in computational efficiency (see Section 2.4), it allows us to easily configure its sensitivity. Many previous works have focused primarily on modelling the distribution of an individual's skin, but we believe that if a process is to be adaptive with respect to the models it generates, it should also be adaptive with respect to the subjects it samples. Being able to configure the detector's sensitivity, therefore, is of great benefit. We have found that with too high a sensitivity, the detector will find the majority of the faces within a given image, but is also liable to return a number of false positive detections, which will harm the accuracy of the model to be generated. Furthermore, with sensitivity set too low, although false positives will be almost entirely eliminated, there is the potential for undersampling, and for a number of subjects to remain undetected, the skin tones of whom might be critical to building an effective colour model. Through experimentation, we believe we have found a sensitivity sweet-spot for our particular process, yielding high precision and moderate recall, but some results have shown this sensitivity to be too high for lower quality or smaller images. This is of no great concern, however, as we run the detection process intelligently, lowering the confidence level required for positive detection until at least one face is found (although stopping short of identifying non-face objects in images that contain no actual faces), all without any significant additional computational overhead.

3.2 Pixel extraction and filtering

Given a set of detected faces (or, a set of regions believed to contain faces), we require methods that can reliably extract the skin pixels that will be used to build our colour model. This is a two-step process. Empirically, we have seen that the detection regions tend to be quite large overestimations of actual face sizes, so we must firstly discard the significant amounts of background information within them. Secondly, we must filter out the non-skin features of faces themselves.

As with face detection sensitivity, we must ensure we are neither significantly oversampling nor undersampling our data. We aim, therefore, to entirely discard the background, whilst retaining as much of the given face as possible. From our experiments, we have found that actual faces can be bounded most reliably using circles, and, hence, we define circular sub-regions within given detection regions, which share the same centres as the original, square detections. We define their radii, however, as being 0.4 times the height of the given detection, granting us an unfiltered skin pixel sample set of ~50% of the original detection pixels for any given image – a larger sample size than any of the previous approaches that we have seen that use the same face detection method, which we believe is advantageous.

The pixels that represent non-skin facial features remain within this set, making it unsuitable for the construction of a skin colour model. Fortunately, however, these non-skin features tend to share one common trait: extreme intensity. Features such as facial hair, nostrils, pupils, and mouths will all usually tend towards black (low intensity), whereas those such as eyes, teeth, and glasses glare usually tend towards white (high intensity). This allows for the filtration of these features to be a relatively trivial process. Despite there being a number of methods to convert colours to greyscale (such as averaging, decomposition, desaturation etc.), we have chosen to calculate luma values, as they have shown to offer the greatest dynamic ranges, affording the greatest potential for simple classifications. For any given pixel *i*, this is achieved using a weighted sum of the RGB component values (designed to approximate the response of human eyes to light intensity), as seen in Eq. 7.

$$luma_i = 0.299 \cdot R_i + 0.587 \cdot G_i + 0.114 \cdot B_i \tag{7}$$

In keeping with the adaptability of the system, we do not predefine acceptable intensity bounds. Instead, we calculate the mean and the standard deviation of the intensities of the pixels in the given extracted set, and use them to establish the range of the filter. Rather than arbitrarily select the number of deviations that defines an outlier, we have studied how closely the varyingly filtered normalised rg clusters match the clusters produced by the manually segmented skin of various images. To accomplish this, we have chosen to compare the standard deviations of the distributions, as this metric gives very strong insight into the shape and size of a given cluster - eminently useful for our purposes of comparison. As to not presume anything about a given cluster's shape, we simply average out the relative differences, in the r and g dimensions, between a given filtered cluster and its respective ground truth cluster, then average these values out across our entire set of test images for each number of deviations the filter is set to allow. The results of this experimentation can be seen in Fig. 3.

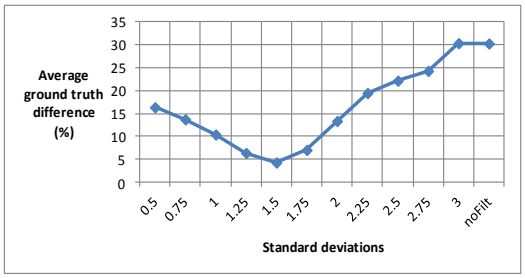


Figure 3: A comparison between the number of deviations that the luma filter allows for and the effect it has on the relative difference in shape between filtered clusters and their respective ground truth clusters.

It should be noted that the only reason we can use such metrics to measure cluster approximation in this instance is because, as we have previously established, skin colour clusters in the normalised rg space tend to be elliptical [20], and allowing for such measurements is another point in favour of using this specific colour space. It should also be noted that we found no correlation between the strictness of the luma filter and the mean r and g values of the filtered sets.

As can be seen from Fig. 3, using 1.5 standard deviations to discard luma value outliers represents a clear optimal choice for yielding accurate filtered pixel sets, as the error is minimised. This number of deviations will retain ~87% of the pixels in the given set, suggesting that an average sub-region will contain ~13% non-skin pixels. This tolerance should prove consistent over any number of input images or face detections, as it only estimates the relationship between face sub-regions and the skin pixels within them, and nothing image-specific. The elimination of such pixels using this process can be illustrated by Fig. 4. After the filtering of all sub-regions within a given image, the remaining pixels will constitute our filtered set.

(a) (b) (c)

Figure 4: (a) The original face detection region, (b) the definition of our sub-region to eliminate background information, and (c) the outlying pixels removed according to their luma values.

3.3 Colour distribution modelling

With a set of pixels that we can be confident are representative of the skin colour distribution within a given image, we are able to build a colour model. There is a wide range of modelling options, but the correct choice for our system can be made with a small number of simple considerations.

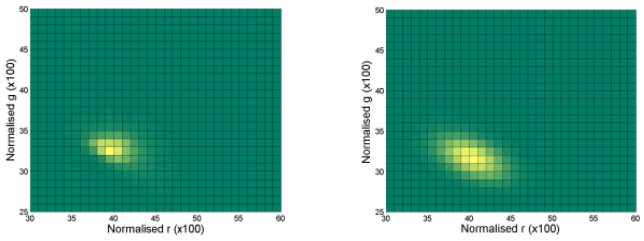
Firstly, we have only positive samples with which to train the model, ruling out any classifier type dependent upon negative samples for training (such as a Bayes classifier). Secondly, naturally, a given filtered pixel set will merely be a subset of all the skin pixels within the given image, so being able to interpolate potentially undersampled (or even entirely unsampled) colours would be ideal. Thirdly, it would be highly beneficial to the accuracy of our system if, instead of establishing independent decision rules, we were to model the relationship between our two dimensions, and use that to classify pixels. We believe this to be true because we reasonably expect clusters to be elliptical, rather than, say, rectangular (which would be the shape of our model if simple upper and lower dimensional bounds were defined). Lastly, given that we have aspirations for our approach to be adopted by real-time systems, the generated models must allow for efficient calculations.

Thankfully, there is one option that meets all of these criteria: a unimodal Gaussian function. The training of such a model is simple, as we need merely calculate the mean and the covariance of our given set of pixels, which can be accomplished through the use of Eq. 8 and Eq. 9, respectively.

$$\mu_s = \frac{1}{n} \sum_{i=1}^{n} c_i \tag{8}$$

$$\Sigma_s = \frac{1}{n-1} \sum_{i=1}^{n} (c_i - \mu_s)(c_i - \mu_s)^T \tag{9}$$

Here, μ_s and Σ_s are the mean vector and covariance matrix of our colour distribution model *s*, respectively, and *n* is the total number of skin colour sample pixels c_i. The mean vector defines the centre point of our cluster within the colour space, and the covariance matrix describes its shape, size, and orientation. Fig. 5 illustrates just how effectively our modelling process, using only the filtered pixels of face detection results, can approximate the colour distribution of the manually segmented skin of a given image.



(a) (b)

Figure 5: (a) The manually segmented skin of Fig. 1(b) in normalised rg space, and (b) our face detection-based colour model for that particular image, approximating the ground truth.

3.4 Skin segmentation

A significant benefit of using a Gaussian function to model a skin colour distribution is that it will allow us to calculate the probability of specific colours representing skin, rather than merely providing us with a binary classification, as typical decision rules would. This makes our approach extremely flexible, as its classification sensitivity can be trivially adjusted to meet specific end-user needs. If, for example, a certain segmentation process necessitates that as much skin as possible be extracted, without paramount concern for false positives, then the probability required for positive classification can be lowered accordingly. Alternatively, certain processes may need to identify pixels with extremely high skin likelihood (such as those involving neighbourhood information), which our approach can also cater for.

Derived from Eq. 6, the probability of a given pixel representing skin can be described by Eq. 10, where *c* is our input colour vector, and μ_s and Σ_s are our trained Gaussian model parameters.

$$p(c|skin) = e^{-\frac{1}{2}(c-\mu_s)^T \Sigma_s^{-1} (c-\mu_s)} \tag{10}$$

Applied across an entire image, we obtain a "skin likelihood image", wherein every pixel is represented by the probability of its original colour being skin. Although useful for evaluation purposes, we do not achieve the

segmentation of any skin without any form of thresholding. Typically, a threshold may define pixels with a skin likelihood of >0.5 as skin (because they are more likely to be skin than not), and those with likelihoods of <0.5 as non-skin, although, as mentioned, thresholds that serve more sophisticated segmentation methods can easily be applied instead. Fig. 6 illustrates the typical appearance of the outputs of our system.

(a) (b) (c)

Figure 6: (a) Our original image, (b) its corresponding skin likelihood image, and (c) the segmentation achieved through the application of a 0.5 probability threshold.

4. RESULTS

To demonstrate the effectiveness of our system, we investigate the quality of its segmentations, and explore its capacity to be used in real-time.

4.1 Skin segmentation analysis

Despite the developmental process of our system being founded on the shortcomings of existing approaches, and our system, therefore, being theoretically capable of yielding better results than the aforementioned, we must test rigorously, and contrast our results with those achieved by others, to prove that this is the case. Unfortunately, there is a distinct lack of common ground when it comes to the datasets used by contributions made to the field of skin segmentation, so we have decided to implement a small number of previous techniques ourselves, and analyse their performances using a subset of images from the DB Skin annotated skin database [26]. This database contains 103 images of varying quality, size, environment, and subject, as well as their respective manual annotations, and, as such, it represents a strong test of various techniques' general-purpose capabilities. The vast majority of the images within this set contain detectable faces, but there are a small number that do not, some of which contain no skin whatsoever. In these latter cases, our system will not build colour models, and will classify entire images as 'non-skin'. The results of our system, therefore, may not compare favourably to those achieved by the static methods for the few images that contain skin but no detectable faces, but this inaccuracy will be more than offset by the results we achieve for the images that contain no skin at all, for which we will have 100% accuracy, and the static methods will often struggle.

We will make objective comparisons to six of the approaches discussed in Section 2. This involves the implementation of the colour space-specific decisions rules defined by Eqs. 1-4, a non-parametric model in the form of a normalised rg Bayes lookup table (constructed through the application of Eq. 5 to 498k skin and 7,029k non-skin arbitrary training pixel-samples), and the face detection-based approach presented by Hsieh et al. [24] (implemented as accurately as possible given the documentation of their system), as it shares similar design philosophies with our approach, and the comparison should, therefore, yield a number of important insights. We believe that using such a broad selection of existing methods to evaluate our own approach will be more than sufficient to establish its efficacy.

In terms of evaluation metrics, we are looking at four different (yet intrinsically related) quantities in the analysis of segmentations, all based upon the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) yielded by the processing of a given image:

- Recall (or Sensitivity) = TP / (TP + FN) (The proportion of actual skin in an image that has been found.)
- Precision = TP / (TP + FP) (The proportion of positive classifications that are actually skin.)
- Specificity = TN / (TN + FP) (The proportion of actual non-skin that has been correctly classified.)
- Accuracy = (TP + TN) / (TP + TN + FP + FN) (The overall image classification accuracy.)

We believe it is important to analyse all of these values, as they will all give different insights into the successes and failings of the methods we are evaluating. The results of testing the various approaches we are interested in over our entire dataset can be seen in Table 1.

Table 1: The results of processing our 103-image dataset using the six previous approaches we have identified, as well as our own system (using a 0.5 probability threshold).

	RGB (Kovac et al. [1])	Norm. rg (Soriano et al. [7])	HSV (Sobottka and Pitas [9])	YCbCr (Hu et al. [15])	Bayes lookup table	Face-based (Hsieh et al. [24])	OUR SYSTEM
Recall	74.19%	43.10%	66.30%	73.70%	44.60%	23.76%	33.29%
Precision	49.88%	41.17%	43.53%	53.37%	46.55%	55.85%	63.28%
Specificity	81.92%	86.39%	80.06%	85.44%	89.54%	89.12%	95.41%
Accuracy	79.65%	79.50%	78.07%	83.05%	82.41%	82.95%	85.55%

4.2 Real-time processing analysis

As well as segmentation accuracy, we must also evaluate our system in terms of its real-time performance. The requirement of computational efficiency has been instrumental in the design of every aspect of our system, including in the choices of face detector, colour space, and modelling method. Therefore, we believe that the benefits to accuracy that our approach is capable of delivering have come at no more of a computational cost than they reasonably could have.

Of course, the definition of "real-time" can vary from application to application, so we present a performance evaluation that considers a potential use-case for our system, and offers frame rate results for a wide range of common image resolutions. In this way, we believe prospective adopters can themselves determine whether our system performs adequately given their specific image size, frame rate, and accuracy requirements. The testing procedure for our investigation was simple, as we set up an input image stream of a single subject interacting with their webcam (as seen in Fig. 7) at a certain resolution, and determined the average number of frames per second our system could process over the course of a minute. This test was carried out for five standard 4:3 image resolutions, using an Intel Core 2 Q9550 2.83GHz processor. The results we achieved can be seen in Table 2.

Figure 7: A sample of the image stream used to investigate our system's real-time capabilities, containing a single subject interacting with a webcam.

Table 2: The results of processing the image stream using our system, with performance represented by the average frame rates achieved at a number of different image resolutions.

Resolution	Pixels	Frames (per sec.)
160 x 120	(19200)	**93.3**
320 x 240	(76800)	**24.2**
480 x 360	(172800)	**9.74**
640 x 480	(307200)	**6.25**
800 x 600	(480000)	**3.7**

5. CONCLUSIONS

In comparison to a wide range of existing segmentation techniques, our system has performed extremely well (see Table 1). The overall accuracy results we have achieved for the DB Skin database are clearly superior to those of the other approaches we have looked at, some of which are very highly regarded in the field. We believe this to be a consequence of our system's flexibility, as the greatly varying input images have been classified using colour models constructed

specifically for them by our system, whereas the static models applied by all but one of the other approaches we have investigated have fallen short. What is interesting about our results is that our system's outstanding accuracy hasn't been caused by any greater skin sensitivity (in fact, the recall rates we achieved using a 0.5 probability threshold were considerably lower than those yielded by most other techniques), but where it most certainly has excelled is in its precision and specificity, and, as such, we can much more confidently claim our positive classifications are actually skin than any other approach can. Should an adopter need levels of sensitivity comparable to the alternative methods, however, our system can certainly deliver them with a trivial adjustment of the applied probability threshold, without too great a detrimental effect on the overall accuracy of results (although with a significant drop in precision).

The results produced by the approach of Hsieh et al. [24] offer additional points of interest for us. Despite our system achieving better results for every investigated metric, the distribution of results across those metrics is very similar, as both approaches achieve comparatively low recall, moderate precision, and high specificity. Given the benefits to accuracy that using adaptive methods can demonstrably yield, the low recall rates with which this is accomplished suggest that the downfall of static methods lies within their overestimation of skin colour distributions, and that greater results can be achieved by focusing on how to more confidently correctly classify non-skin, rather than through the pursuit of greater sensitivity.

In terms of real-time performance, we believe our system has proven to be capable of producing results at a rate that would be adequate for a number of potential applications (see Table 2). In our testing, we have used a setup that resembles common hand detection-based user interactivity, for which large resolutions are unnecessary, and achieved, at the more relevant lower resolutions, frame rates that should prove to be more than sufficient for most hand detection methods and applications. It should be noted that the results we have presented have been achieved without the implementation of any application-specific performance-enhancing modifications, such as the frame-by-frame region-of-interest definitions utilised to great effect by Hsieh et al. [24] for hand detection purposes, greatly reducing the amount of time spent looking for a face given its previous location and the amount of time spent detecting skin given the face's current location. Although the implementation of such enhancements would be simple, and the benefits to results significant, we have decided to present our system in an entirely generalised form, as to not limit the scope of its flexibility.

Despite being extremely pleased with our system's performance, there are a number of extensions we would like to investigate in the future. For example, as with the vast majority of skin segmentation techniques, our system does struggle to adequately deal with specular reflections on skin. This is not a simple issue to overcome, as specular reflections will typically tend towards pure white, which is a colour not expected to be skin under any normal circumstances, and, in fact, is a colour that we attempt to filter out of our pixel sets before building our models. We believe that making use of pixel neighbourhood information could lead to significant progress in this regard, but any benefits would naturally come at a frame rate cost. Additionally, we will consider the adoption of a non-face detection-based skin colour classifier for our system, to be used during instances of no faces being detected within given frames. Although this would improve the general segmentation performance of our system, we would have to reconcile this concept with the fact that our approach has been designed to exclusively operate on images that contain detectable faces, and the application of it to a problem that doesn't guarantee their presence would be somewhat redundant.

REFERENCES

[1] Kovac, J., Peer, P. and Solina, F., "Human Skin Colour Clustering for Face Detection," Proc. EUROCON 2003 – Int. Conf. Computer as a Tool 2, 144-148 (2003).

[2] Vezhnevets, V., Sazonov, V. and Andreeva, A., "A Survey on Pixel-Based Skin Color Detection Techniques," Proc. GraphiCon 2003, 85-92 (2003).

[3] Fleck, M., Forsyth, D. A. and Bregler, C., "Finding naked people," Proc. European Conf. Computer Vision 1996 2, 593-602 (1996).

[4] Brand, J. and Mason, J., "A comparative assessment of three approaches to pixel-level human skin-detection," Proc. Int. Conf. Pattern Recognition 2000 1, 1056-1059 (2000).

[5] Zarit, B. D., Super, B. J. and Quek, F. K. H., "Comparison of Five Color Models in Skin Pixel Classification," Proc. Int. Workshop Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems 1999, 58-63 (1999).

[6] Oliver, N., Pentland, A. P. And Berard, F., "LAFTER: Lips and Face Real Time Tracker," 1997 IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 123-129 (1997).

[7] Soriano, M., Martinkauppi, B., Huovinen, S. and Laaksonen, M., "Skin detection in video under changing illumination conditions," Proc. 15th Int. Conf. Pattern Recognition 1, 839-842 (2000).

[8] Störring, M., Andersen, H. J. and Granum, E., "Physics-based modelling of human skin colour under mixed illuminants," Robotics and Autonomous Systems 35, 131-142 (2001).

[9] Sobottka, K. and Pitas, I., "A novel method for automatic face segmentation, facial feature extraction and tracking," Signal Processing: Image Communication 12(3), 263-281 (1998).

[10] Phung, S. L., Bouzerdoum, A. and Chai, D., "Skin segmentation using color pixel classification: analysis and comparison," IEEE Trans. Pattern Analysis and Machine Intelligence 27(1), 148-154 (2005).

[11] Jordao, L., Perrone, M., Costeira, J. P. and Santos-Victor, J., "Active Face and Feature Tracking," Proc. Int. Conf. Image Analysis and Processing 1999, 572-576 (1999).

[12] Berbar, M. A., Kelash, H. M. and Kandeel, A. A., "Faces and Facial Features Detection in Color Images," Proc. Geometric Modeling and Imaging – New Trends, 209-214 (2006).

[13] Ahlberg, J., "A System for Face Localization and Facial Feature Extraction," Tech. Rep. LiTH-ISY-R-2172 (1999).

[14] Tathe, S. V. and Narote, P. S., "Face detection using color models," World J. Science and Technology 2012 2(4), 182-185 (2012).

[15] Hu, M., Worrall, S., Sadka A. H. and Kondoz, A. M., "A Fast and Efficient Chin Detection Method for 2-D Scalable Face Model Design," Int. Conf. Visual Information Engineering 2003, 121-124 (2003).

[16] Elgammal, A., Muang, C. And Hu, D., "Skin Detection – a Short Tutorial," [Encyclopedia of Biometrics], Springer-Verlag, Berlin & Heidelberg, 1218-1224 (2009).

[17] Jones, M. J. and Rehg J. M., "Statistical Color Models with Application to Skin Detection," Int. J. Computer Vision 46(1), 81-96 (2002).

[18] Yang, M.-H. and Ahuja, N., "Gaussian Mixture Model for Human Skin Color and Its Applications in Image and Video Databases," Proc. SPIE 3635: Storage and Retrieval for Image and Video Databases VII, 458-466 (1999).

[19] Lee, J. Y. and Yoo, S. I., "An Elliptical Boundary Model for Skin Color Detection," Proc. Int. Conf. Image Science, Systems, and Technology 2002, 472-479 (2002).

[20] Terrillon, J.-C. and Akamatsu, S., "Comparative Performance of Different Chrominance Spaces for Color Segmentation and Detection of Human Faces in Complex Scene Images," Proc. 12th Conf. Vision Interface 2, 180-187 (2000).

[21] Viola, P. and Jones, M. J., "Robust Real-Time Face Detection," Int. J. Computer Vision 57(2), 137-154 (2004).

[22] Mittal, A., Zisserman, A. and Torr, P. H. S., "Hand detection using multiple proposals," Proc. 22nd British Machine Vision Conf., 75.1-75.11 (2011).

[23] Wimmer, M. and Radig, B., "Adaptive Skin Color Classificator," ICGST Int. J. Graphics, Vision and Image Processing 6(Special Issue on Biometrics), 41-46 (2006).

[24] Hsieh, C.-C., Liou, D.-H. and Lai, W.-R., "Enhanced Face-Based Adaptive Skin Color Model," J. Applied Science and Engineering 15(2), 167-176 (2012).

[25] Liao, W.-H. and Chi, Y.-H., "Estimation of Skin Color Range Using Achromatic Features," Proc. Eighth Int. Conf. Intelligent Systems Design and Applications 2, 493-497 (2008).

[26] Ruiz-del-Solar, J. and Verschae, R., "Skin Detection using Neighborhood Information," Proc. 6th Int. Conf. Automatic Face and Gesture Recognition, 463-468 (2004).