A SOM Based Approach to Skin Detection with Application in Real Time Systems

David A Brown[1,2] Ian Craw[1] Julian Lewthwaite[2]

[1]Mathematical Sciences
University of Aberdeen
http://www.maths.abdn.ac.uk

[2]AXEON Ltd
Davidson House, Aberdeen
http://www.axeon.com

Abstract

A large body of human image processing techniques use skin detection as a first primitive for subsequent feature extraction. Well established methods of colour modelling, such as histograms and Gaussian mixture models have enabled the construction of suitably accurate skin detectors. However such techniques are not ideal for use in adaptive real time environments.

We describe methods of skin detection using a Self-Organising Map or SOM, and show performance comparable (94% accuracy on facial images) to conventional techniques. We also introduce the AXEON Learning Processor as the basis for a hardware skin detector, and outline the potential benefits of using this system in a demanding environment, such as filtering Internet traffic, to which conventional techniques are not best suited.

1 Introduction

Skin is arguably the most widely used primitive in human image processing research, with applications ranging from face detection and person tracking to pornography filtering. Skin detection techniques can be both simple and accurate, and so can be found in many commercial applications, for example the driver eye tracker developed by Ford UK [13].

At its simplest, skin pixel detection can be based on (say) the claim that a pixel of colour (r, g, b) is skin when $r > g$ or $r > b$, or perhaps both. Such a criterion shows some independence to the overall illumination level, and can be applied very rapidly; however its limitations are obvious in all but the most controlled environments, where the colour red can appear on non-skin objects.

More elaborate skin detectors can be constructed if we have a collection of labelled pixels. A conceptually simple, if resource intensive method works with a collection of (say) 256^3 bins, one for each possible colour value[1]. Each skin pixel from the training set is added to the bin corresponding to its own colour; the resulting bin counts then give the observed frequency of occurrence of that colour amongst the training pixels. It is natural at this stage to consider this as a probability distribution: a threshold is chosen and a pixel is identified as skin if the probability of its colour occurring exceeds this threshold. The threshold can be set so that 95% of skin pixels are correctly identified, and the detector assessed by measuring false acceptance rates with a non-skin set [2].

[1]It has been shown [6] that effective generalisation requires a much smaller histogram e.g. 32 bins per colour channel.

BMVC 2001 doi:10.5244/C.15.51

The resulting distribution is large, and depends on the choice of training set, so it is natural to try to model its essential features. It has been demonstrated [12] that the skin colour distribution can be well modelled by a mixture of Gaussians, giving a very much more compact description with essentially no effect on the ability to distinguish skin pixels in test, as opposed to training, situations. This method has shown reasonable success in practise; McKenna *et al* [9] demonstrate a tracking system based on modelling per object colour distributions able to work at standard frame rates.

Although the Gaussian mixture model provides a very succinct representation of the distribution explicit in the histogram, it too has a number of problems. The Expectation-Maximisation (EM) algorithm [10] used to fit a mixture model to training data requires that the number of distributions used be specified *a priori*; in skin modelling this is not known accurately and is dependent on the colour model used. Techniques for estimating the number of clusters (*model order*) exist [1] but there is no universal agreement on which method is most effective. The EM algorithm also takes no account of the cyclic nature of some components of popular colour spaces, for example hue in hue-saturation-value, and often further coordinate transforms or translations must be performed for best results. Re-parameterising or adapting a mixture model is also often required in object tracking applications, and is a complex and intensive task.

There is thus reason to seek an alternative way of gathering and approximating the skin distribution, which requires less data pre-processing and provides a less intensive update procedure while still retaining the compact nature of the mixture model. The Self-Organising Map (SOM) [7] opens the possibility of skin detection as effective as that by a mixture model, but without the practical difficulties.

This paper investigates two ways in which a SOM can be used to detect skin. The results in Section 4 suggest there is little to choose between the mixture model and SOM in terms of accuracy of recognition; as such the practical simplicity, ability to retrain rapidly and hardware implementation provide significant benefits in certain applications.

Sections 2 and 3 briefly introduce the SOM and describe how it can be used for skin detection. Section 4 provides a performance comparison against the mixture model for facial skin detection and details performance on the Compaq data set [6]. An outline of a hardware implementation using the AXEON Learning Processor is presented in Section 5, and we draw conclusions in Section 6.

2 The Self Organising Map

Devised by Kohonen in the early 80's, the SOM is now one of the most popular and widely used types of unsupervised artificial neural network. It has mainly been used to find patterns in and classify high dimensional data, although it works equally as well with low dimensional data. The basic SOM consists of a 2-dimensional lattice L of neurons. Each neuron $n_i \in L$ has an associated codebook vector $\boldsymbol{\mu}_i \in \mathbb{R}^n$. In what follows $n = 2$, although in other applications n is often much larger. The lattice is either rectangular or hexagonal (see Figure 1), with the connections within L determining the neighbourhoods of a given neuron.

Training the SOM involves first randomly initialising all the codebook vectors and then sequentially presenting each training sample[2]. We first fix a metric on L, usually

[2] Other training algorithms exist, see [7] for details of the batch algorithm.

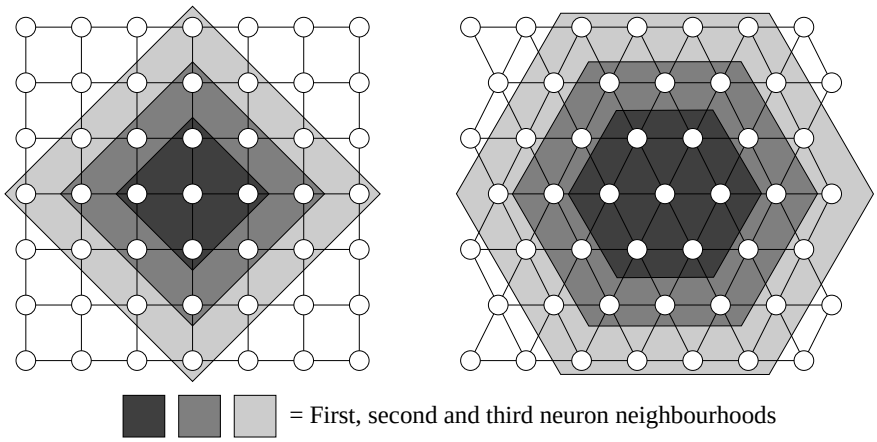

Figure 1: The SOM Lattice. Lattices are either rectangular or hexagonal; this in turn determines how many neurons lie in each neighbourhood. All skin extraction experiments have used a hexagonal lattice.

Euclidean or Manhattan. Each $\boldsymbol{\nu} \in \mathbb{R}^n$ is presented as an input vector to all neurons in the network, and the winning neuron n_c with codebook vector $\boldsymbol{\mu}_c$ is determined so that

$$||\boldsymbol{\nu} - \boldsymbol{\mu}_c|| = \min_{1 \leq i \leq k} ||\boldsymbol{\nu} - \boldsymbol{\mu}_i||$$

where k is the number of neurons in the network. The neurons in a specific neighbourhood of the winning neuron then have their codebook vectors adjusted to be closer to the input vector according to a parameterised learning function. As training progresses, the learning rate and the size of the affected neighbourhood are decreased, and the lattice gradually forms a topologically ordered mapping (or feature map) of the training data.

If necessary, a calibration phase then takes place, where labelled training data is sequentially presented to the SOM and, each time, the data label and index of the winning neuron recorded. Each neuron is then assigned the label of the type for which it 'fired' the most. Classification can then take place by presenting data and labelling with the label of the winning neuron each time.

The SOM package used in these tests is the freely available SOM_PAK C implementation [8].

3 Approaches to Skin Detection

The first choice which must be made when building any skin detector is which colour space to use. It is generally agreed that intensity has a large effect on (r, g, b) triples and that it is desirable to remove an intensity component before subsequent analysis, but there is much less agreement on how the remaining two colour parameters should be defined. We therefore take an inclusive approach, and have performed tests using the following four colour spaces, each of which has previously given good results.

Hue, Saturation and Value (HSV) Essentially a description of colour space in cylindrical polars, the HSV colour space forms a hexacone, with black at the main vertex

and white at the centre of the base. The central axis of the hexacone gives the V co-ordinate, while (H, S) is a polar description of a point on the 'colour wheel'.

Cartesian Hue-Saturation (XY) A different representation of the above using Cartesian rather than polar co-ordinates to describe the colour wheel:

$$X = S\cos H, \qquad Y = S\sin H.$$

Tint, Saturation and Luminance (TSL) A more complex alternative to HSV proposed and shown to be viable in [12].

Normalised RG A colour model commonly used in face detection; it reduces the sensitivity to illumination changes while staying very close to the 'usual' (r, g, b):

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

We have assumed no knowledge of camera properties throughout, and so have avoided CIE variant and sRGB colour spaces. One interpretation of some recent work [12] is that there is little benefit to be gained from camera knowledge in the type of unconstrained skin detection we consider here.

In each case we can now represent 'intensity invariant' colour as a 2-dimensional vector $[c_1, c_2]^T$. For testing consistency we assume that each component is equally likely to affect the chance of a pixel being skin, and homogeneously rescale each colour component so $c_1, c_2 \in [0, 1]$. These colour vectors can now be used to train a SOM.

The first of our SOM-based detectors requires training solely on skin pixels. We obtain skin pixels from landmarked face images as indicated in Figure 2. The pixels within the skin areas are formed into a randomly ordered list and sequentially presented to the SOM as described in Section 2. Following training, classification takes place by measuring the quantisation error, taken as the Euclidean distance between the codebook vector of the winning neuron and the given sample. Because the trained SOM is essentially a feature map of the training data, skin samples are likely to have a smaller quantisation error than non-skin samples. A threshold for classification is chosen and each new pixel is declared to be skin if its quantisation error is lower than this threshold.

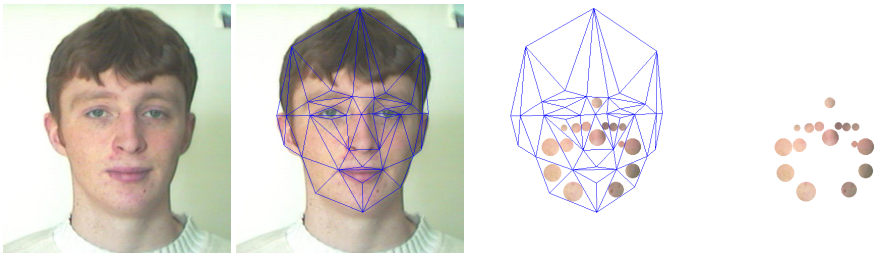

Figure 2: Skin pixel sampling. Areas within certain landmark triangles are assumed to contain skin.

Our second SOM-based skin detectors additionally requires a sample of non-skin pixels. These were taken from a random sample of Internet images, carefully filtered to

remove images containing humans. The SOM is then trained on the randomly ordered list of skin and non-skin pixels. This list is also used to calibrate the SOM, labelling the neurons as either skin or non-skin. Classification takes place as detailed in Section 2.

For our experiments, we have used a list of 30000 skin pixels to train both the skin-only SOM and also a Gaussian mixture model. Details of the mixture model used can be found in [1]. A list containing 15000 skin pixels and a second list of 15000 non-skin pixels were used to train the skin/non-skin SOM.

4 Test Strategy and Performance Analysis

We have been primarily concerned with detecting faces, and detecting skin in face images, and so have used two independent face imagesets as sources for skin samples, both of which are avaliable with manually located landmarks. The first, the Aberdeen Pilot (AP) [3] consists of 396 frontal face images giving over 2 million skin samples using the extraction method summarised in Figure 2. The second, the Illumination Capture (IC) imageset, consists of 160 frontal face images under varied lighting conditions giving around 750000 skin samples. We have also used a sample of 4 million non-skin pixels taken from a set of 113 images collected through random web browsing. We discarded pixels of very high and low intensity due to unstable hue and saturation readings.

Each detector was trained separately using a filtered random sample from each distinct imageset, and then tested using all the filtered samples from all 3 imagesets. For the SOM detectors we experimented with between 16 and 256 neurons using a hexagonal lattice. We varied the thresholds both for the skin-only SOM and mixture model and each individual test with a specific training set, SOM size and threshold was repeated for every colour model. We measured correct detection rates for each of the 3 imagesets and ranked the configurations by best average percentage. The tests were carried out on a Sun Ultra 60, and both the SOM_PAK and Cluster packages were used without modification.

In general there was little difference in overall performance between detectors. We tested over 300 detector configurations in all, and at best each detector was able to achieve an overall average of over 94%. There were however some important differences which became apparent when we looked at the top 20-30 results for each detector.

Both SOM detectors were able to achieve comparable results with all colour spaces, but the mixture model showed consistently better results with the normalised RG colour space. The plots in Figure 3 of the AP training data perhaps gives an insight into this behaviour, showing the normalised RG representation to be arguably the most tightly clustered. It is difficult to conclude that this is the only explanation; the SOM detectors however clearly show no such colour model 'preference'.

Although there was little difference between top performances, the skin only SOM was observed to be the most consistent detector. Both the mixture model and the skin/non-skin SOM showed a 5% performance loss over their respective top 20 configurations, while the skin-only SOM showed a drop of less than 1% over its top 30 configurations. This observation suggests that choice of configuration is more critical to the effectiveness of a mixture model than a SOM. Such robust behaviour is desirable, especially in more unconstrained situations such as skin pixel detection in Internet images. Where there is little knowledge of the content of images, it is difficult to tweak and test different colour models and parameters to improve results and so a detector which can offer near optimal

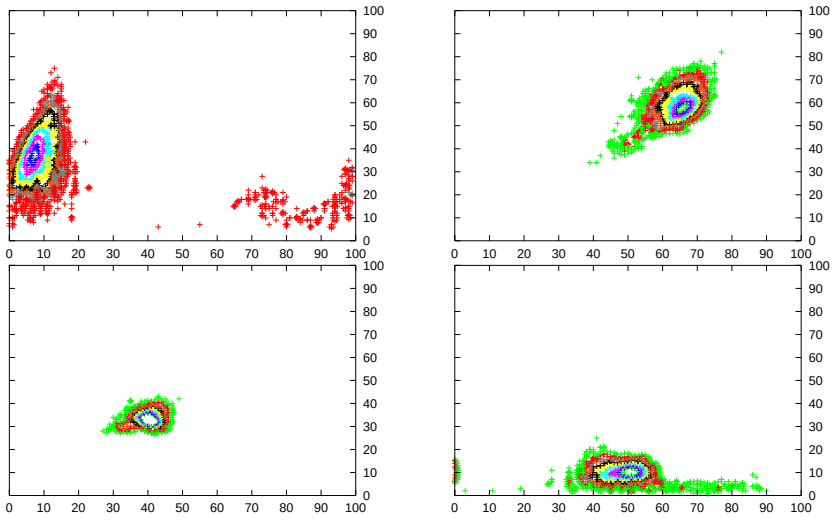

Figure 3: Cumulative histograms of the 30000 Aberdeen Pilot training pixels in the four different colour spaces tested. The clusters in some colour spaces can be more easily modelling with a Gaussian mixture model, or even a single Gaussian.

results with only an educated guess at configuration is potentially attractive.

It is also worth noting that the skin-only SOM performed marginally (1-2%) but consistently better than the mixture model in detecting skin pixels from the same imageset as the training set. One reason for this behaviour is the inherent limitation of a mixture model to approximation using a linear combination of 'elliptical clusters'. The HS plot in Figure 3 provides a good example of a distribution where using a number of overlapping Gaussians could introduce approximation errors not present in the SOM, due to the unconstrained flexibility of the lattice. However we only offer this as an intuitive explanation; the SOM still resists rigourous mathematical analysis, a discussion of the relative arguments can be found in [7].

Although the skin/non-skin SOM provided the best result of all our experiments (95.5%), it was generally less consistent than the skin-only SOM, However, testing still produced some useful points. Figure 4 shows the u-matrix [5] of the SOM configuration which gave the best result.

The u-matrix demonstrates graphically the SOM's ability to identify clusters in data; there is a definite area of neurons representing skin pixels bordering a larger area of loosely clustered non-skin neurons. Certain colour models and SOM sizes showed a number of neurons which appeared to be in the wrong place or have the wrong label, caused by the overlap between skin and non-skin distributions. Although such occurences can be problematic, their overall number can be used to give an idea of the overlap between the distributions; such measures have previously proved useful in choosing which colour model to use [12].

Despite the obvious theoretical and practical difficulties with non-skin samples, we have been able to construct a very simple but effective skin detector very easily. Other simple statistical detectors, including histograms, have tended to be much larger and more rigid; the u-matrix highlights the SOM as small, threshold free, easily adaptive and also

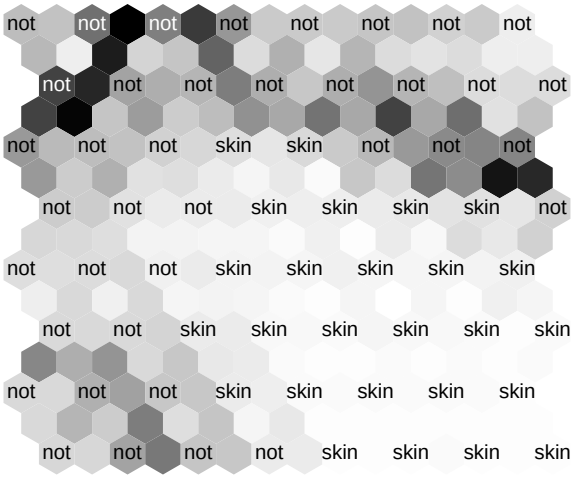

Figure 4: The skin/non-skin u-matrix. The u-matix is a commonly used graphical representation of a SOM. Every second (labelled) hexagon represents a neuron in the SOM, and the interlacing (un-labelled) hexagons are shaded to represent the distance between codebook vectors of adjacent neurons - the lighter the hexagon, the closer the codebook vectors. This u-matrix represents a 64 neuron SOM trained on 15000 Illumination Capture skin pixels and 15000 non-skin pixels from random Web images using the normalised RG colour space. It shows the tight clustering of skin neurons, with a clear border of relatively unclustered non-skin neurons.

suitable for hardware implementation as discussed in Section 5.

Both SOM detectors performed well with a variety of different SOM sizes; between 64 and 196 neurons giving good results with all colour models. This is quite a wide range, and although configuration is not critical with the skin-only SOM, it would be desirable to narrow it with further experimentation, perhaps looking for trends towards optimal sizes. This could potentially be done with a growing SOM [4], where the number of neurons is dependent on quantisation error and variable during training. This growing behaviour is not supported by the SOM_PAK.

Although facial skin detection was our primary aim, we have also performed tests using the Compaq image database [6], which consists of over 13000 random Web images with marked up skin pixels, with no buttons or icons. This provided over 80 million labelled skin pixels and over 800 million non-skin pixels. We increased the number of random training samples to 200000 (0.25% of the total Compaq skin samples) to cater for the much wider colour spread in the Compaq set, and trained a skin-only SOM with 64 neurons using the TS colour space (a randomly chosen configuration from the skin-only SOM top 30). Testing on the full Compaq set with a threshold of 0.1 we achieved accuracy of 78% on skin pixels and 68% on non-skin pixels. While this is a step down from Jones' histogram result [6] of 88% equal error rate, it is encouraging given the use of less training data, arbitrary choice of configuration and less resource. We expect that a larger training sample and larger SOM sizes (256+ neurons) would give improved results; more neurons would better cover the wider colour spread.

Figure 5: Sample facial skin detection. Skin detection has been performed using a 64 neuron skin-only SOM trained on 30000 Aberdeen Pilot normalised RG skin samples with a threshold of 0.02. None of these test images belong to the training set. The results are fairly mixed - definite success on images with prominent faces, but failure on more complex scenes. We also note here that the training set contains mostly Caucasian subjects.

Figure 5 shows some sample detection results. These samples are typical of our wider results on individual images - showing reliable facial skin detection on images with prominent faces, but fairly mixed results with more multi-object scenes. Many of these raw results could be improved with post-processing techniques such as shape analysis and/or edge-based segmentation, but occurences of, for example, skin coloured buildings could still be problematic. It is also worthwhile noting that while in theory it could be argued that the facial skin colour distribution is representative of full body skin colour, this is *not* the case with unconstrained Internet imagery. The wider the spread of the Compaq skin sample and our experimental results clearly support this argument.

5 Hardware Implementation

One distinct benefit of a SOM based approach to skin detection is the availability of SOM hardware. Discussion of some particular hardware solutions is provided by Kohonen [7]; we however will focus on the AXEON Learning Processor, a parallel processor with a neuron array capable of 1 million classifications per second with only moderate clock speeds. It is designed so it can be cheaply incorporated into most mainstream workstations and servers.

The combination of fast versatile hardware with robust SOM skin detection has potential application in high speed or batch image processing systems, where conventional techniques may introduce undesirable loads or bottlenecks. Although both histograms and mixture models have proved reasonably effective in colour based image segmentation and object tracking [9, 11], the very nature of the training and adaption process makes them unsuitable for particular real time situations. Thus, for example, pornography filtering software on a central web gateway or proxy could easily introduce further processor load, impairing connection handling performance. The hardware solution is also favourable in minimalist embedded applications such as CCTV auto-tracking and in-car eye detection and monitoring. The SOM uses less resource than the frequently used histogram approach while additionally being easily adaptable in such rapidly changing environments.

Work is currently taking place to exploit the power of the Learning Processor in the Web environment, with investigations into other applicable core image processing techniques, such as feature detection, taking place.

6 Conclusions

We have introduced the SOM as an effective statistical model for skin colour distributions, and used the SOM to construct skin tone detectors comparable to those built with conventional histogram and mixture model techniques. Our experiments on facial skin have achieved consistent accuracy of over 94%.

The bijective equivalences between different colour spaces do *not* preserve Gaussians; as such the assumption that skin colours can be modelled as a mixture of Gaussians is a different assumption in different colour spaces. Our experiments confirm that this matters in practice. By contrast the SOM is theoretically 'coordinate free' over a wide range of choice of input, and as such we avoid the problem of making a choice of colour representation. Again these results have confirmed this to some extent in practice. We have also eliminated the problem of model order selection, and have replaced it with the less critical problem of deciding SOM size.

As well as detecting facial skin, we have shown good results in the more demanding web based detection environment. This has motivated hardware implementation with the Learning Processor, which along with the SOMs practicality and adaptability has created a core colour image processing component with some advantages over currently favoured methods.

In conclusion, we have explored the SOM as a skin detection tool with practical advantages over both mixture models and histograms in certain circumstances. Our future work aims to build on these core results, and extend the theory into actual applications which exploit these advantages.

References

[1] C. A. Bouman. *Cluster: An Unsupervised Algorithm for Modelling Gaussian Mixtures*. School of Electrical Engineering, Purdue University, http://dynamo.ecn.purdue.edu/~bouman/software/cluster, September 2000.

[2] J. Brand and J. S. Mason. A Comparative Assessment of Three Approaches to Pixel-level Human Skin-Detection. In *Proc of the International Conference on Pattern Recognition*, volume 1, pages 1056–1059, 2000.

[3] I. Craw, N. P. Costen, T. Kato, and S. Akamatsu. How should we represent faces for automatic recognition? *IEEE:T-PAMI*, 21(8):725–736, 1999.

[4] M. Dittenbach, D. Merkl, and A. Rauber. The Growing Hierarchical Self-Organizing Map. In *Proc of the International Joint Conference on Neural Networks (IJCNN 2000)*, volume VI, pages 15 – 19, Como, Italy, 2000.

[5] J. Iivarinen, T. Kohonen, J Kangas, and S. Kaski. Visualising the clusters on the Self-Organizing Map. In *Multilple Paradigms for Artificial Intelligence (SteP94)*, pages 122–126. Finish Artificial Intelligence Society, 1994.

[6] M. M. Jones and J. M. Rehg. Statistical Color Models with Application to Skin Detection. In *Proc of the Conference on Computer Vision and Pattern Recognition*, volume 2, pages 274 – 280, Fort Collins, Colorado, 1999.

[7] T. Kohonen. *Self-Organizing Maps*. Information Sciences. Springer, third edition, 2000.

[8] T. Kohonen, J. Hymminen, J. Kangras, and J. Laaksonan. SOM_PAK: The Self-Organizing Map program package. Technical Report A31, Helsinki University of Technology, 1996.

[9] S. McKenna, Y. Raja, and S. Gong. Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, 17:225–231, 1999.

[10] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26:195–239, 1984.

[11] K. Schwerdt and J. L. Crowly. Robust Face Tracking using Color. In *Proc of the International Conference on Face and Gesture Recognition*, pages 90 – 95, Grenoble, France, 2000.

[12] J.-C. Terrillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative Performance of Different Skin Chrominance Models and Chrominance Spaces for the Automatic Detection of Human Faces in Color Images. In *Proc of the International Conference on Face and Gesture Recognition*, pages 54 – 61, Grenoble, France, 2000.

[13] D. Tock and I. Craw. Tracking and measuring drivers' eyes. *Image and Vision Computing*, 14:541–548, 1996.