Skin Patch Detection in Real-World Images

Hannes Kruppa, Martin A. Bauer and Bernt Schiele

Perceptual Computing and Computer Vision Group
ETH Zurich, Switzerland
http://www.vision.ethz.ch/pccv
email: {kruppa, bauerm, schiele}@inf.ethz.ch

Abstract. While human skin is relatively easy to detect in controlled environments, detection in uncontrolled settings such as in consumer digital photographs is generally hard. Algorithms need to robustly deal with variations in lighting, color resolution, and imaging noise. This paper proposes a simple generative skin patch model combining shape and color information. The model is parametric and represents the spatial arrangement of skin pixels as compact elliptical regions. Its parameters are estimated by maximizing the mutual information between the model-generated skin pixel distribution and the distribution of skin color as observed in the image. The core of this work is an empirical evaluation on a database of 653 consumer digital photographs. In addition, we investigate the potential of combining our skin detector with state-of-the-art appearance-based face detectors.

1 Introduction and Related Work

Skin detection plays an important role for example in tracking people, in filtering out adult web images, or in facilitating human-computer interaction. We are especially interested in skin detection as a cue for detecting people in real-world photographs. The main challenge is to make skin detection robust to the large variations in appearance that can occur. Skin appearance changes in color and shape and is often affected by occlusion (clothing, hair, eye glasses etc.). Moreover, changes in intensity, color and location of light sources affect skin appearance. Other objects within the scene may cast shadows or reflect additional light and so forth. Imaging noise can appear as speckles of skin-like color. Finally, there are many other objects in the world which are easily confused with skin: certain types of wood, copper, sand as well as clothes often have skin-like colors.

Physics-based approaches to skin color modeling [12] use spectrographic analysis to derive a physical reflectance model of skin. Skin reflectance is usually described by its thin surface layer, the epidermis, and a thicker layer underneath, the dermis. The light absorption in the dermis is mainly due to the ingredients in the blood such as haemoglobin, bilirubin and beta-carotene which are basically the same for all skin types. However, skin color is mainly determined by the epidermis transmittance which depends on the *dopa-melanin* concentration and hence varies among human races [12]. Skin color appearance can then be represented by using this model and by incorporating camera and light source parameters. In uncontrolled scenes, however, these parameters are not known.

Its rotation and scale invariance make skin color especially useful for real-time tracking systems. However, gesture trackers for human-computer interaction rely on controlled lighting conditions [11]. In other scenarios like outdoor surveillance, potential illumination changes have to be addressed. Recently, approaches have been proposed which automatically adapt skin color model parameters by analyzing color differences of consecutive images [10]. Assuming motion continuity these approaches can deal with gradual changes in ambient light. They do not apply to still images though.

One of the most comprehensive accounts on skin color models for uncontrolled still images is due to Jones and Rehg [6]. In their case, a skin color model is learned from a huge collection of web images. A Bayesian classifier for skin color is then constructed which also incorporates a model of the non-skin class. The approach relies on color alone. Fleck and Forsyth [4] and Wang et al [13] propose systems for filtering adult images by finding naked people. In the approach by Fleck and Forsyth a combination of low-level image filters is used combining skin color and texture features.

In this paper, we introduce a generative skin patch model combining color and shape information (section 2) and present results of a large empirical evaluation (section 3). As today's state-of-the-art face detectors do not make use of skin concepts, we also investigate the potential of combining skin and face detection in section 4. Finally, section 5 draws conclusions and outlines possible directions for future research.

2 Approach: A Generative Skin Patch Model

Rather than relying on skin color alone the proposed approach combines color information with shape constraints to locate skin patches. Allowable shapes of skin are embodied by a generative skin patch model. The shape parameters are estimated by maximizing the mutual information between the model-generated skin pixel distribution and the distribution of skin color as observed in the image. In the current implementation the skin patch model is represented as an unrotated ellipse with state variables $\gamma = (x_c, y_c, w, h)$ where (x_c, y_c) denotes the center and (w, h) the dimensions. Ellipses as shape primitives are frequently used for modeling the human body, in particular the head, arms and limbs [15]. The shape model is denoted by S and is employed by the algorithm for generating a distribution $p(\boldsymbol{x} = skin|\gamma)$, which for each image location $\boldsymbol{x} = (x, y)$ represents the probability that the corresponding pixel belongs to a skin patch. The model $p(\boldsymbol{x} = skin|\gamma)$ is represented by a piecewise constant function

$$S(\gamma) = S(x_c, y_c, w, h) = \begin{cases} \frac{1}{1+exp^{-a}} & : \quad \frac{(x-x_c)^2}{w^2} \pm \frac{(y-y_c)^2}{h^2} \leq 1 \\ 0 & : \quad else \end{cases}$$

where the parameter a in the logistic function (c.f. [1]) is increased towards the boundary to smooth out probabilities. Thus the proposed generative model embodies two intuitive properties about the spatial distribution of skin color: First, skin is distinguished as a contiguous region and second, skin often appears in oval shapes. Restricting the model to an unrotated oval introduces a bias for facial skin.

The core idea of the algorithm is to derive the parameters of S from a complementary cue, a skin color model C, thus combining shape and color information. We

employ a Gaussian chrominance model with parameters θ which is trained from data using Maximum Likelihood Estimation. This color model has been shown to work for different skin complexions of the different human races [5]. At this stage one can also consider the use of color constancy techniques such as [3]. This implies the standard trade-off between discrimination and invariance. Even though not reported here our experience suggests that simple color models are well suited to the task especially when enough training data is available. Similar observations have been reported in [6]. For aligning shape and color information we now maximize the mutual information between $p(\boldsymbol{x} = skin|\gamma)$ and $p(\boldsymbol{x} = skin|\theta)$ searching the parameter space of γ:

$$\arg\max_{\gamma} I(S(\gamma), C(\theta)) \tag{1}$$

Maximizing mutual information is an alignment technique which maximizes statistical dependence. The concept has been used successfully in several vision and machine learning tasks e.g. for feature selection [2], for audio-visual data association [8] as well as for robust registration [14] which probably comes closest to the way it is used in this paper. In the following two section we will present qualitative and quantitative evidence that this alignment technique is robust to noise and discontinuities typical of real-world imaging. There is a direct relationship between mutual information and the Kullback-Leibler (KL) divergence. The KL-divergence between a probability mass function $p(u, v)$ and a distinct probability mass function $q(u, v)$ is defined as:

$$D(p(u, v)||q(u, v)) = \sum_{u_i, y_v} p(u_i, v_j) \cdot \log \frac{p(u_i, v_j)}{q(u_i, v_j)} \tag{2}$$

This relative entropy or information divergence measure is commonly used as a pseudo distance between two distributions. By defining $q(u, v) = p(u) \cdot p(v)$ the mutual information can be written as the KL-divergence between $p(u, v)$ and $p(u) \cdot p(v)$:

$$I(U; V) = D(p(u, v)||p(u) \cdot p(v)) \tag{3}$$

Mutual information therefore measures the "distance" between the joint distribution $p(u, v)$ and the product distribution $q(u, v) = p(u) \cdot p(v)$, which are identical if and only if they are independent.

Instead of resorting to stochastic parameter sampling as in [14] we derive a deterministic form of gradient ascent in mutual information space for efficiently estimating γ: The parameter space is traversed using an adaptive local grid which is succinctly centered over maxima of $p(\boldsymbol{x} = skin|\theta)$ starting at the global maximum of this distribution and continuing at other maxima in descending order. At each iteration the algorithm follows the steepest gradient adapting the center point (x_c, y_c) or the dimensions (w, h). Typically, convergence is reached in about 5 to 10 iterations. Once the algorithm has converged γ represents a single computed skin region hypothesis.

For generating multiple skin patch hypotheses the skin distribution $p(\boldsymbol{x} = skin|\theta)$ is then reshaped. More specifically, after a hypothesis has been formed the associated region is first excluded from the original distribution and after, the scheme is repeated. The value of mutual information is used to decide if a hypothesis is valid and if the search is to be continued. After a predefined number of examined hypotheses with a low mutual information value, the algorithm stops.

3 Experiments: Skin Detection

To evaluate the performance of the proposed skin detector we first examine retrieval and precision rates *on the pixel level* similar to [6]. In particular, the performance is compared to the naive approach using only color.

The whole test database contains 653 color JPEG images. For skin detection these have been downsampled from a 3.3 megapixel resolution to 150 by 100 pixels. From a randomly chosen subset of 30 images, 53978 skin pixels where labeled manually. The fotos cover a wide range of real-world situations both indoor and outdoor (meetings, parties, ski-trips, beach scenes etc.). Figure 1 shows several representative example im-

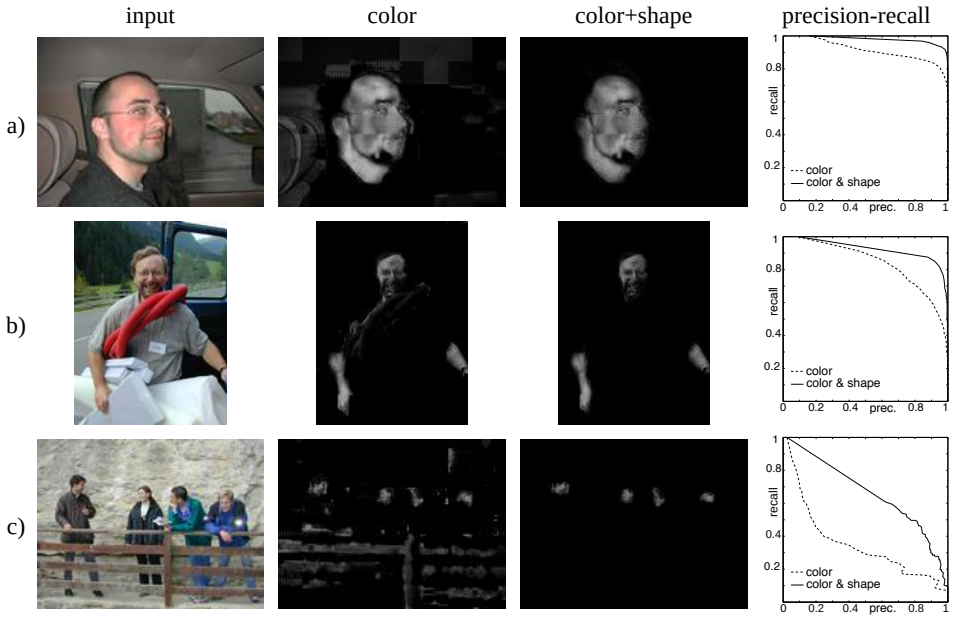

Fig. 1. The first two columns show input images and corresponding skin color distributions. Column three shows the hypotheses generated by the proposed skin detector combining color and shape information. The right column shows precision-recall curves, comparing the purely color based approach to the proposed scheme. See text for a detailed discussion of these examples.

ages. Input images are shown in the first column while the second column shows the distribution of skin color $p(\boldsymbol{x} = skin|\theta)$. Column three shows the output of the proposed skin detector. Here only those color probabilities are shown which have been classified by the proposed detector as being part of a skin region. The last column in these figures shows precision-recall curves of skin pixel detection as a function over a detection threshold. Two curves are shown for each image: one is based on evaluating color alone (dotted line) the other (solid line) plots the results of the proposed detector.

Figure 1a) shows many false positives in color (car interior) as well as JPEG-artefacts which appear as block structures in the upper part of the image. As can be seen here, lossy compression used in image formats like JPEG cause discontinuities in $p(\boldsymbol{x} = skin|\theta)$. Note that a connected-component approach would have difficulties to separate the full head from the car interior. With the proposed approach the equal error rate is improved from 85% to 95% in this example.

Figure 1b) shows a 15% increase in equal error rate. In this image the shape model separates out the skin portions from the red-colored hose which is wrapped around the person's shoulder. Note again that a connected-component approach would have difficulties to separate the hose from the person's face. The equal error rate is improved here from 70% to 85%. The advantages of the proposed detector become most evident as the amount of skin-like color in the image increases. In example 1c) the detector improves the equal error rate by 25%. In this image a wooden fence causes numerous false positive detections. Wood is a well-known distractor [6] which frequently occurs in both indoor and outdoor situations. The wooden fence and all remaining false positives are removed by the detector. Altogether, an equal error rate of 60% is reached while the color-based approach attains only 35%. A few false negatives occur from people's hands in the image because they hold directly on the wooden bars which makes hands and wood indiscernible for the algorithm.

While figure 1 shows only a small selection for lack of space, these images are representative for the obtained results on all images[1]. These results clearly demonstrate the advantage of integrating color and shape information. The shape constraints embodied by the skin patch model successfully eliminate a substantial amount of false positives which leads to improved detection performance. In particular, the detector proves to work robustly in unconstrained real-world photographs.

4 Experiments: Skin Detection vs. Face Detection

Unlike face recognition (i.e. subject identification), *face detection* is a classification problem concerned with locating all faces within arbitrary backgrounds. An extensive survey on face detection with more than 150 references appeared only recently [16]. Two state-of-the-art face detectors are due to Rowley et al [7] and Schneiderman et al [9]. Both approaches are appearance-based and only use intensity information. Rowley's neural network approach can handle in-plane rotations of faces but not out-of-plane rotations (profiles). However, profiles can be expected to occur much more often in photographs. To our knowledge Schneiderman's Bayesian face classifier is still one of the best systems in terms of detection performance. Since it is still the only system to support profile views it may even be *the* best face detector available today. Since both systems model only specific portions of the appearance space they are vulnerable to the many variations that occur in the real-world. Attempts to model the complete appearance space can be expected to fail because of degrading discriminance.

Although a face can be regarded as a special case of a skin patch, neither of the two face detectors makes use of skin concepts. Yet, from an algorithmic viewpoint,

[1] Results on the full data are available on the web (URL blinded to preserve anonimity of the authors)

the proposed skin detector has at least three substantial advantages over face detectors: it is faster, it works at smaller resolutions and, most importantly, it is more robust to variation in appearance. In addition, this section also presents empirical evidence that the skin detector is robust to lighting changes.

Appearance based face detectors need to search all image locations at multiple scales. In contrast, the skin detector examines only skin colored regions at one scale which results in significantly faster execution. It also requires less image support allowing to detect very small skin patches. Figure 2 shows a few examples showing the outputs of the two face detectors and the proposed skin detector. Row 2(a) illustrates the effect of occlusion and rotational variation. Both face detectors fail on the right face since they model the appearance of a complete face only. Rowley's approach misses the other face, too, because the particular facial appearance is not part of the model. Both faces are detected by the skin-based approach. Since the proposed scheme allows for discontinuities, it also works when people wear beards, piercings or glasses. Face-like

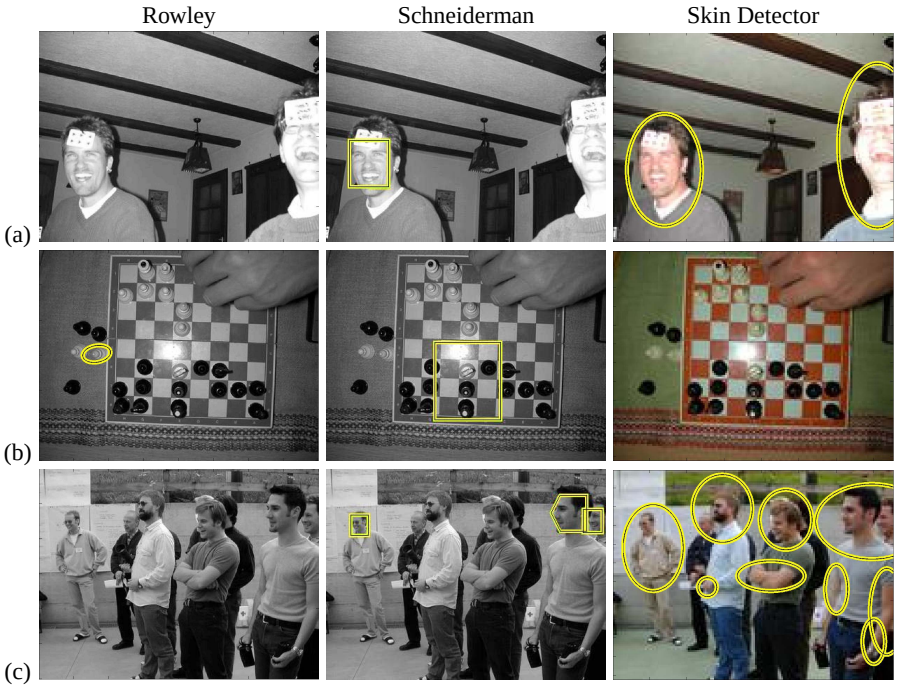

Fig. 2. Characterization of appearance-based face detection vs. skin detection. The first two columns show the face detection results of Rowley and Schneiderman, row three shows the skin finder's output. These examples illustrate characteristic effects of (a) occlusion and facial expression, (b) face-like distractors, (c) crowds

distractors pose additional challenges. For instance, checkered surfaces are likely to be

	true positive	false negative	false positive	precision	recall
Schneiderman	387	305	165	70.1%	55.9%
skin detector OR Schneiderman	639	53	1458	30.5%	92.3%
skin detector AND Schneiderman	263	429	12	95.6%	38.0%
Rowley	150	542	94	60.5%	27.7%
skin detector OR Rowley	562	130	1397	40.2%	81.2%
skin detector AND Rowley	103	589	2	98.1%	14.9%

Table 1. A quantitative account of appearance-based face detection (Schneiderman, Rowley) and its combination with the proposed skin detector. Results are from a test set of 653 real-world consumer photographs containing 2243 faces. Here, Schneiderman's scheme compares favorably to Rowley's. When combining Schneiderman's face detector with the proposed skin finder, recall (OR-combination) or precision (AND-combination) is leveraged to above 90% in both cases.

confused with facial features like mouth, nose and eyes. This often occurs with shirts and other clothes, in figure 2(b) it occurs with a chess board. Note that in this example, the skin detector misses the chess player's hand. This is because the current implementation only allows for unrotated elliptical shapes. Row 2(c) shows a more complex scene with a crowd of people. Faces are in profile view, some of them are only partially visible. Rowley has no detections, Schneiderman returns three faces. The skin detector retrieves all six faces and some additional skin regions.

Next we quantitatively compared the performance of the combined scheme to the individual face detectors on all 653 images containing 2243 faces. For evaluating Schneiderman's approach we uploaded our database to the on-line demo[2]. Rowley provided a copy of his system. The results are very promising. The skin finder returned 74.4% of all faces, whereas Schneiderman's face detector has a recall rate of 55.9% and Rowley's scheme 27.7%. That is, the skin detector's recall rate in detecting faces is almost 20% higher than Schneiderman's algorithm and 47% higher than Rowley's approach. As can be expected we found complementary results for precision. Since the proposed skin detector is designed to return skin region in general, not just faces, its precision is only 28.3%. Rowley's scheme reaches 60.5% and Schneiderman 70.1% on this data set. As performance of skin and face detectors turned out to be *complementary* we examined their combinations. When counting all faces found by either Rowley's scheme OR the skin detector the recall rate is boosted to 81.2% (versus an initial rate of 27.7%). Precision is raised to 98.1% (versus 60.5%) when counting only those faces found by both detectors. Results from combining the skin detector with Schneiderman's approach are equally encouraging: Precision is as high as 95.6% (versus 70.1%) which is comparable to the combined performance using Rowley's approach. Recall is raised to 92.3% using a logical OR combination. This is even higher than the combination with Rowley's scheme. These results indicate that the combination of skin and face concepts can lead to a substantially better face detection performance. Depending on the type of combination Schneiderman's face detector in combination with the proposed skin finder reaches precision or recall rates above 90%.

[2] http://vasc.ri.cmu.edu/cgi-bin/demos/findface.cgi

5 Conclusion and Future Work

This paper proposes a skin patch detector integrating color and shape information. An empirical evaluation of the detector on real-world photographs yields two main results: First, there is a clear benefit in modeling skin as approximately contiguous regions of certain colors *and* shapes rather than relying on color alone. In particular, the proposed detector proves to work robustly in unconstrained real-world photographs. Second, appearance-based face detectors should be combined with skin detection for their *complementary* strengths and weaknesses. In future work we aim to analyze skin specific specularities and to encode this information within the generative model.

References

1. Christopher M. Bishop, editor. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
2. A. Colmenzrez and T. Huang. Face detection with informationbased maximum discrimination. In *CVPR*, pages 782–787, 1997., 1997.
3. G.D. Finlayson, B.V. Funt, and K. Barnard. Color constancy under varying illumination. In *ICCV'95*, pages 720–725, 1995.
4. Margaret M. Fleck, David A. Forsyth, and Chris Bregler. Finding naked people. In *ECCV (2)*, pages 593–602, 1996.
5. Hideo Fukamachi Jean-Christophe Terrillon, Mahdad Shirazi and Shigeru Akamatsu. Skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, March 2000.
6. Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. In *CVPR*, pages 274–280, 1999.
7. H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *PAMI*, 20(1):23–38, 1998.
8. D. Roy and A. Pentland. Learning words from natural audio-visual input. In *International Conference of Spoken Language Processing*, December 1998.
9. H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*, June 2000.
10. Leonid Sigal and Stan Sclaroff. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. In *CVPR*, 2000.
11. T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *SCV95*, page 5B Systems and Applications, 1995.
12. Moritz Stoerring, Hans J. Andersen, and Erik Granum. Skin colour detection under changing lighting conditions. In *7th International Symposium on Intelligent Robotic Systems'99*, pages 187–195, July 1999.
13. James Ze Wang, Jia Li, Gio Wiederhold, and Oscar Firschein. System for screening objectionable images using daubechies' wavelets. In *International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services*, pages 20–30, 1997.
14. W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi–modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–52, march 1996.
15. Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19(7):780–785, 1997.
16. Ming-Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *PAMI*, 24(1):34–58, January 2002.