What impacts skin color in digital photos?

Albrecht Lindner and Stefan Winkler

Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, Singapore

ABSTRACT

Skin colors are important for a broad range of imaging applications to assure quality and naturalness. We discuss the impact of various metadata on skin colors in images, i.e. how does the presence of a metadata attribute influence the expected skin color distribution for a given image. For this purpose we employ a statistical framework to automatically build color models from image datasets crawled from the web. We assess both technical and semantic metadata and show that semantic metadata has a more significant impact. This suggests that semantic metadata holds important cues for processing of skin colors. Further we demonstrate that the refined skin color models from our automatic framework improve the accuracy of skin detection.

Keywords: skin color, color model, memory color, image semantics, statistical test, image-mining

1. INTRODUCTION

Skin colors are an important topic in image processing and are used for various tasks such as face detection/recognition, tracking of body parts, or color correction. However, skin colors in images are influenced by many different factors ranging from the pictured person (e.g. skin type, degree of tanning) over surrounding factors (e.g. illumination, geographic location) to technical factors (e.g. camera type, flash setting). In this paper we pursue a holistic approach to assess the impact of different factors on skin colors in images. We present a statistical framework that estimates skin colors fully automatically on a large database of images and provides significance scores to judge the quality of the estimation. We then use the framework to assess and discuss the impact of different factors on skin colors in images. We further show that skin detection is improved by using the refined skin color models from our framework.

A large-scale analysis of skin color requires a large database of images showing human faces and skin. We use three sources to acquire images for this study from the internet: Google image search, Flickr, and the publicly available PubFig[1] dataset. Google provides a specific search option for images of type *face* that can be accessed through the advanced search options or its API.[2] Flickr does not provide this option, but it is possible to increase the chance of downloading face images by using keywords that dominantly occur with face images. In the first part of this publication (Section 3) we explain how we use Google and Flickr to acquire face datasets related to specific metadata attributes such as keywords or camera type. We want to highlight that our study is based on images that have been rendered in-camera or even post-processed by the photographer before uploading to the world wide web.

In Section 4 we describe a statistical framework to analyze the downloaded images, where each image is described by a three-dimensional histogram in CIELAB color space. We use the Mann-Whitney-Wilcoxon test to assess whether certain histogram bins have a significantly higher or lower bin count in face images in comparison to non-face images. This results in a significance distribution in color space that culminates in the bin with the most prominent color in face images. A complete assessment of the entire distribution indicates which colors are likely to be skin color and which are not. We run a different test for each metadata attribute to acquire a specific significance distribution per attribute.

Section 5 presents the results and discusses the impact of different technical and semantic metadata on skin colors. We show that skin colors are strongly impacted by the depicted person and by associated keywords. The impact is weaker for geographic location, camera type or flash setting.

Finally we outline in Section 6 how the estimated skin color models can be used to improve skin detection, and we demonstrate at the example of a simple algorithm that a specific skin color model outperforms a general skin color model. This shows that metadata is an important aspect to improve skin color processing in digital images.

A. Lindner is now with Qualcomm, San Diego. E-mail: alindner@qti.qualcomm.com.

2. RELATED WORK

Skin color is a memory color, which means that humas are able to recall the color because it is very familiar to us. The invention of the Munsell Color System[3] gave rise to systematic research on memory colors. An early publication from Bartleson provides measurements for *flesh*, *tanned flesh*, and eight other memory colors in the Munsell color space.[4]

Memory colors are important for image quality because deviated memory colors make images appear unnatural.[5] This is the reason why fine-tuning memory colors is common practice for skin tones[6,7] as well as for other memory colors.[8] It is debatable which memory color is the most important, but we think that skin color does play a dominant role especially for the consumer market. We thus focus on skin colors for the rest of this article.

Skin color models are the basis of many algorithms such as skin segmentation,[9,10] face detection,[11,12] hand detection,[13] or color correction.[14] It is obvious that a better skin color model helps to improve such algorithms because relevant regions of an image can be found with higher precision. This paper presents methods that yield more precise skin color models and therefore is fundamental to a variety of subsequent higher-level imaging applications.

Our approach of computing skin colors from web images is similar in spirit to the work from Jones and Rehg.[15] However, we use a statistical significance test instead of a simple histogram division which can lead to undesired effects if the bin count in the denominator is close to zero. The significance test additionally provides a mathematical measure for the quality of the estimation. We also do multiple estimations on different datasets in order to assess which type of metadata has stronger impact on skin colors in images.

There are numerous factors that influence skin colors such as ethnic group or age.[16] Nevertheless we focus on factors that are commonly found in metadata of photographic images and that are machine readable. This covers technical metadata, i.e. EXIF file header, as well as semantic metadata, i.e. associated keywords or user profiles in online communities.

3. DATASETS

We acquire large datasets of images in order to reliably determine skin color distributions with statistical methods. Depending on the metadata we use different sources for the acquisition. This section describes the datasets used in this paper.

3.1 Frequently used keywords in face images

In a first step we want to determine a list of keywords that frequently occur in images with faces because we use these keywords later to query for images. For this purpose we use the MIRFlickr database with 1 Million annotated images.[17] We use the Viola and Jones face detector[18] that is part of the OpenCV package,[19] which detects faces in 189,224 images. We then create a keyword histogram from all images with a detected face.

The frequency counts in this histogram are naturally biased by the overall frequency of a keyword. One possibility to compensate for this is to divide by the keyword frequency in the entire MIRFlickr database. This however can over-emphasize very rare keywords, e.g. for a keyword that occurs in only one image that happens to contain a face, the normalization by simple division leads to a ratio of infinity. We thus use the Kullback Leibler divergence to estimate a keyword's dominance in face images with respect to all images:

$$D_{KL}(p,q) = p \cdot \log\left(\frac{p}{q}\right), \tag{1}$$

where p and q are the probabilities that a keyword occurs in the face image subset or in the complete dataset, respectively. The 10 keywords with largest divergence are: *portrait* (37.2), *people* (16.6), *girl* (16.5), *bw** (15.8), *woman* (12.4), *2008* (12.0), *nikon* (11.7), *film* (10.1), *street* (9.9), *me* (9.6).

*abbreviation: *black and white*

3.2 Technical metadata: EXIF

We use Flickr's public API[20] to download images with technical metadata. It allows to query for images with specific keywords, geolocation or other attributes. In addition to the image, it is possible to download the image's EXIF header, keywords, comments from the Flickr community, and other related data.

To increase the probability that the downloaded images contain human faces, we focus the search on images annotated with the top 30 keywords with highest Kullback Leibler divergence D_{KL}. To avoid biased images we remove keywords related to colors (e.g. *bw*), specific persons (e.g. *obama*), geographic locations (e.g. *japan*) or a specific year (e.g. *2008*). The following independent datasets have been downloaded:

- Geogrid: We overlay a 6×12 grid over the globe with cell sizes of 30 degrees in latitude and longitude directions, as shown in Figure 2. We download up to 1000 images for each keyword and grid cell, resulting in a total of 68,795 images.
- Cameratype: Flickr does not allow to query images of a given camera type, but we can exclusively download images that define the camera type in the EXIF header and record this information together with the image. In total we obtain 25,585 from various camera manufacturers and camera types. The two most frequent manufacturers are Canon (10,232 images) and Nikon (8,210 images).
- Flash: Some cameras also record the flash settings; we download 22,387 images that provide this information. Examples of flash settings are *"On, Red-eye reduction"*, *"Auto, Fired"*, or *"Off, Did not fire"*. In total there are 20,398 images without and 1,989 image with flash.

3.3 Semantic metadata: keywords and PubFig

In addition to the technical metadata we also use datasets with semantic metadata:

- Keyword: We use Google's custom search API[2] to download images related to specific keywords. The search can be confined to websites from a specific country or in a specific language using the country and language restrict fields, respectively. To ensure we only acquire face images we use Google's image type field and set it to *face*. The keyword dataset contains up to 200 images per keyword from different Western and Asian countries. We queried for keywords such as *beauty*, *light-skinned*, or *sunburn* for each country separately. The keywords were translated by native speakers to the respective languages, i.e. 美女 as the Chinese translation of *beauty*. This dataset contains 8,585 images covering 47 keywords in 8 languages.
- PubFig: This is a publicly available database of face images depicting different celebrities.[1] We downloaded all images from the publicly available URL list that were still available: 24,713 images from 140 different persons. We do not use the Labeled Faces in the Wild dataset as it contains fewer images per person and thus makes a statistically significant evaluation more difficult.

3.4 Non-skin images

The statistical analysis (see next section) requires a set of negative images, i.e. images that do not contain faces or human skin in general. For this purpose we use images from the MIRFLICKR-25000 image collection.[21] We use the 30 keywords with highest D_{KL} divergence from the previous analysis and discard all images that are annotated with at least one of them. Further we remove all images where the OpenCV[19] face detector triggered. This pruning might not be sufficient to remove all images that show skin colors, but it is good enough to achieve statistically significant results as shown in the following sections.

4. STATISTICAL ANALYSIS

The statistical analysis compares descriptors from two different sets of images. The first set contains face images that all have one specific metadata attribute A in common. This can be a flash setting, a geolocation, a keyword or others. The second set contains images without faces or skin regions and is the same set for all statistical tests throughout this article (see Section 3.4). We refer to these sets as $\mathbf{I}_A$ and $\mathbf{I}_\ominus$, respectively.

Each image in I_A and $I_\ominus$ is described by its color histogram characteristic. We assume that the images are encoded in sRGB and convert them to CIELAB because it is a perceptual color space. We use a color histogram in CIELAB color space and focus on the color region in which skin colors can occur; we choose the intervals $L \in [0,\ 100]$ for the lightness channel and $a, b \in [-15,\ 60]$ for the chromatic channels with 31 equidistant bins along each dimension. To make the statistical analysis more robust we use the OpenCV face detector[19] and compute the histogram characteristic only within the detected areas of the images. Images in which the OpenCV detector did not detect any faces are discarded. Characteristics of the negative set $\mathbf{I}_\ominus$ are computed globally. We refer to the two sets of characteristics as $\mathbf{C}_A$ and $\mathbf{C}_\ominus$, where A is the metadata attribute that images in $\mathbf{I}_A$ have in common.

The statistical analysis is based on the Mann-Whitney-Wilcoxon significance test, which assesses whether the values in one set are significantly larger or smaller than the values in a second set.[22] To compute the test statistic the values of the united set $\mathbf{C}_A \cup \mathbf{C}_\ominus$ are sorted in increasing order. Then the positional indexes of $\mathbf{C}_A$'s values are summed up resulting in the test statistic, which is also referred to as the ranksum T. For example, let $\mathbf{C}_A = \{2, 0.5, 2.2\}$ and $\mathbf{C}_\ominus = \{1.6, 2.3, 4, 3.5\}$, then the sorted unified set is $\overset{1}{0.5}, \overset{2}{1.6}, \overset{3}{2}, \overset{4}{2.2}, \overset{5}{2.3}, \overset{6}{3.5}, \overset{7}{4}$ with the positional indexes stacked on top and the ranksum is $T = 1 + 3 + 4 = 8$.

Under the null hypothesis – both sets contain equally large values – the expected mean and variance of the test statistic and the normalized z score can be computed as follows:

$$\mu_T = \frac{N_A(N_A + N_\ominus + 1)}{2}, \qquad \sigma_T^2 = \frac{N_A N_\ominus (N_A + N_\ominus + 1)}{12}, \qquad z = \frac{T - \mu_T}{\sigma_T}, \tag{2}$$

where N_A and $N_\ominus$ are the cardinalities of the sets $\mathbf{I}_A$ and $\mathbf{I}_\ominus$, respectively.

Larger cardinalities usually entail larger z scores because more samples generally increase the confidence of a test's result. This is unpractical if test results from different tests with different cardinalities have to be compared. We therefore normalize the z score to a reference cardinality N_A^* that yields $z^* \approx \sqrt{N_A^*/N_A} \cdot z$ under the condition that $N_\ominus \gg N_A$.[23] We use the standardized z^* score in this publication for an easier comparison of test results. More examples for other characteristics can be found in Lindner.[23]

In contrast to usual hypothesis testing we do not use a significance threshold to confirm or reject the null hypothesis. Instead, we use the significance score as a measure of relatedness between a characteristic, e.g. a color value, and a metadata attribute.

5. RESULTS FOR AUTOMATIC SKIN COLOR ESTIMATION

Figure 1 illustrates the output of the statistical significance test, where the first set contains all face images of all downloaded datasets. The left *ab*-plane shows a cross section of the significance distribution at the lightness level where it achieves its maximum. The contours indicate equidistant levels from the distribution's maximum to zero (outermost blue line). The distribution's maximum is marked with a cross. The column plot on the right shows the same contours along the lightness axis through the distribution's maximum. The distribution's maximum is located at an average flesh tone, and that the distribution reaches further into the reddish regions than towards the neutral gray axis.

5.1 Technical metadata

Figure 2 shows skin color estimates for different geographic locations from the geogrid dataset. Each grid cell's color is defined by the maximum significance value of its images' significance distribution. The thin white lines depict country borders for a better orientation. Grid cells with fewer than 10 face images are black. Skin colors in images from the regions of Africa and India are darker than from other regions, even though the effect is not very strong. It is worth noting that the skin color of the respective Flickr users themselves might be as important as the images' geographic locations. This is obvious if a Flickr user only takes images of himself no matter where on the planet he is.

Figures 3(a) and 3(b) show significance distributions for Canon and Nikon cameras from the cameratype dataset. Skin colors of Nikon cameras are slightly more saturated and are more confined along the L-axis. The

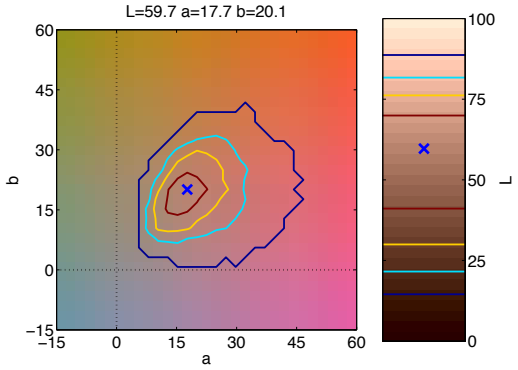


Figure 1. Contour plot of the significance distribution derived from the statistical framework, based on all face images of all datasets. The three-dimensional distribution is depicted with contour lines on the *ab*-plane and along the lightness axis. The blue cross is located at the distribution's maximum, whose CIELAB coordinates are indicated in the title. The contour lines visualize the significance distribution in equidistant steps from the maximum value to zero (outermost blue contour line).

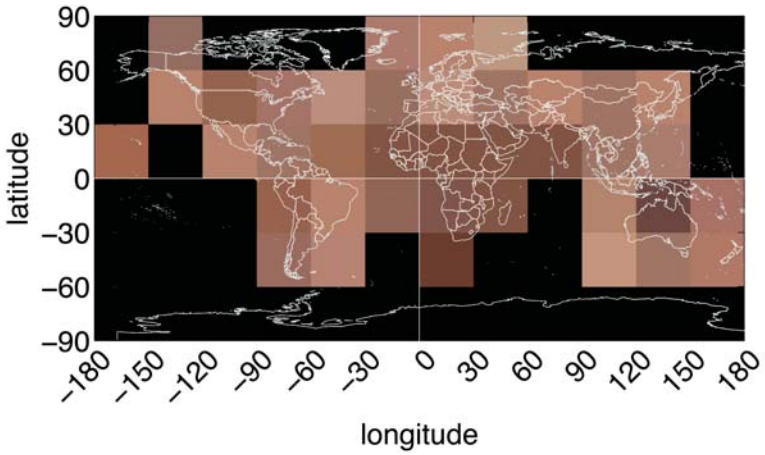
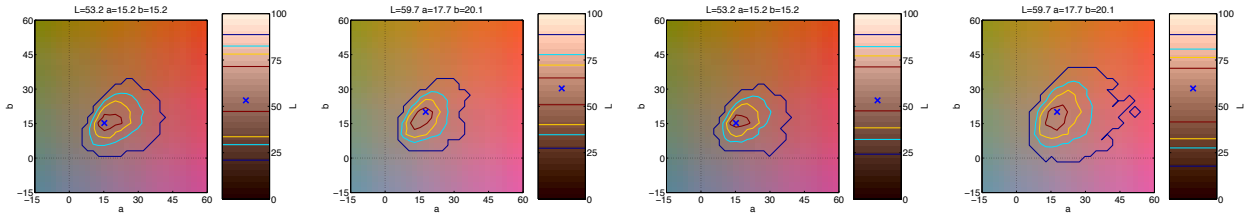


Figure 2. Estimated skin colors for cells of a geographic grid. Country borders are outlined with thin white lines for better orientation. Grid cells with fewer than 10 images are in black due to insufficient significance.

plots in Figures 3(c) and 3(d) show contour plots of the significance distributions of images taken with and without flash, respectively. The skin color distribution in images taken with flash is more elongated and reaches slightly higher reddish saturations in comparison to images taken without flash.

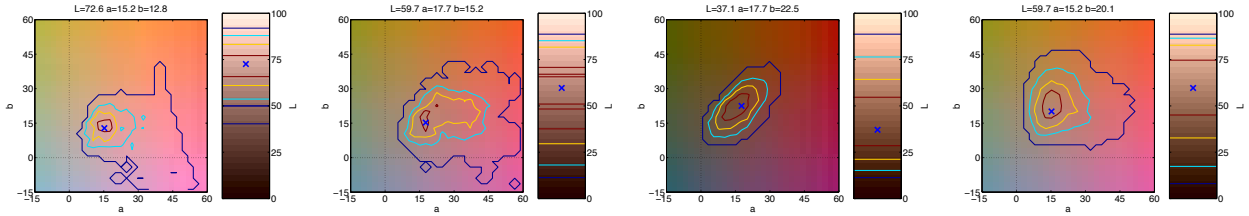


(a) Canon (b) Nikon (c) Without flash (d) With flash

Figure 3. (a, b) Significance distribution for skin colors of Canon and Nikon cameras, respectively. (c, d) Plots showing the distributions for images taken without and with flash, respectively.

5.2 Keywords and PubFig

Figures 4(a) and 4(b) show distributions for the keywords *geisha* (芸者)[†] and *sunburn* for images downloaded from Japanese and Canadian websites, respectively. Figures 4(c) and 4(d) exemplarily show the skin color distributions for two persons from the PubFig database: *Barack Obama* and *Nicole Kidman*. We see again how the automatically estimated significance distributions precisely reflect the properties of the underlying metadata attribute associated with the images used for the analysis.



(a) Keyword: *geisha* (jap) (b) Keyword: *sunburn* (can) (c) PubFig: *Barack Obama* (d) PubFig: *Nicole Kidman*

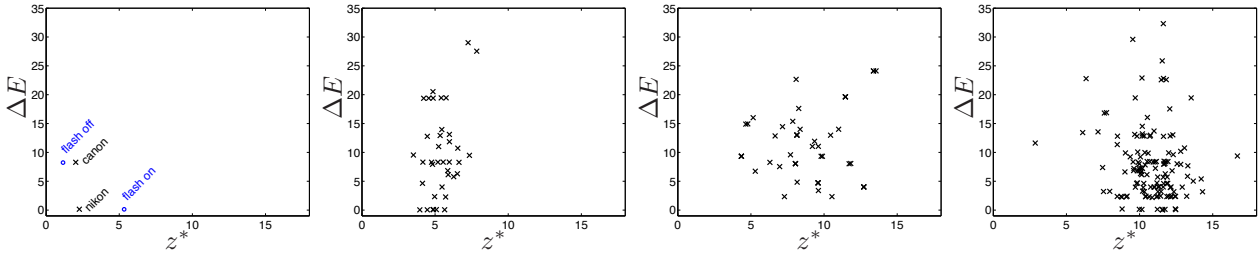
Figure 4. (a, b) Skin color distributions for two examples from the keyword dataset: *geisha* (Japan) and *sunburn* (Canada). (c, d) Skin color distributions for two examples from the PubFig dataset: *Barack Obama* and *Nicole Kidman*.

5.3 Significance analysis

The significance scores z^* indicate how strongly a metadata attribute impacts an image's skin color. If the impact is low, then the variety of skin colors is high, and thus the distribution is widespread and has a low maximum. On the other hand, if a metadata attribute has a strong and unique impact, the distribution is narrow and has a higher peak value.

Figure 5 shows scatter plots of the z^* distribution's maximum value and the ΔE distance to the overall skin color estimated from the entire dataset (see Fig. 1). For all datasets, the technical metadata has comparatively low significance values. The flash setting and camera type do not even have a strong impact on the distribution's peak location. In contrast, semantic metadata shows much higher significance values, especially for different persons from the PubFig dataset. The keywords with highest significance are *beauty* and *headache*[‡] downloaded from different anglo-saxon countries.

This shows that semantic metadata is a much richer source of information to estimate expected skin color distributions in images than technical metadata. We emphasize again that our downloaded images have been rendered in-camera or even post-processed by the photographer before uploading to the web. If the same statistical analysis were done with raw images, the technical metadata might have a much stronger impact.



(a) camera and flash (b) geogrid (c) keyword (d) PubFig

Figure 5. Scatter plots for the different datasets indicating the significance scores z^* and the ΔE distances to the mean skin color from Figure 1. Technical metadata (camera type and flash setting) and geographic location have comparatively lower significance scores, indicating a weaker impact on skin colors. Semantic metadata such as keywords and depicted person have a stronger significance, suggesting that this type of metadata provides a more reliable clue about the expected skin color distribution in images.

[†]Geishas are Japanese women with prominent make-up to whiten their skin.

[‡]These images show very pale faces.

6. APPLICATION TO SKIN DETECTION

Detecting skin region in images is an important pre-processing step for skin-tone enhancements or other optimization tasks for images depicting people. We thus want to evaluate the use of our refined skin color distributions to detect skin regions. We use the keywords *sunburn*, *paleness* and *beauty* for images downloaded from the United States. The keywords were chosen to reflect a skin color deviation along the chroma axes, along the lightness axis, and a more abstract keyword, respectively. We randomly select 50 images for each keyword and manually create a binary ground truth for all skin regions with one pixel precision. We then run the statistical framework on the remaining images for all three keywords to create new color models.

The detection we employ uses a pixel-by-pixel approach where we convert each pixel's sRGB values to CIELAB and look up the significance value in the appropriate significance distribution. If the pixel's significance value is above a threshold, the pixel is classified as skin and non-skin otherwise.

The images of each of the three keywords are used twice; once with the keyword specific significance distribution, and once with the significance distribution of all images from the United States. We repeat all experiments for increasing threshold values from the distribution's minimum to maximum value. Figure 6 shows the resulting precision-recall curves for these 6 experiments. We see that the keyword specific distributions improve the precision and recall values over almost the entire threshold range.

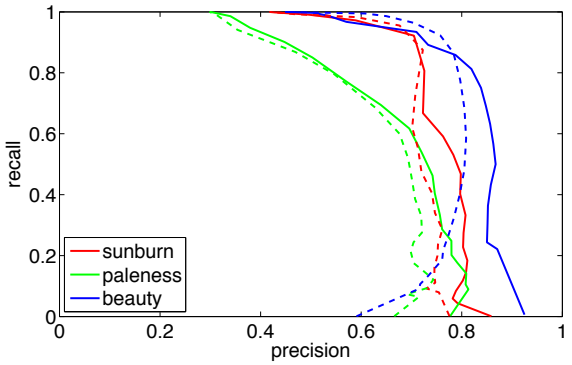


Figure 6. Precision-recall curves for skin detection. There are two curves for each of the three keywords: the solid line is for detections that use the keyword specific color model, whereas the dashed line is for detections with a general color model. The skin detection is better if the keyword is taken into account.

7. CONCLUSIONS AND FUTURE WORK

This paper focused on skin color in images and how it is influenced by different metadata attributes. We investigated both technical and semantic metadata and collected appropriate image datasets from the web using Google Image Search, Flickr, PubFig,[1] and MIRFlickr.[17] The datasets are listed in Section 3.

We estimated skin color models for all metadata attributes as shown in Figures 2–4 using a fully automatic statistical framework as explained in Section 4. The framework not only provides the color model, but also significance scores that indicate the quality of the estimated model. An analysis of the significance scores for the different datasets shows that semantic metadata has a comparatively stronger impact on skin colors than technical metadata as visualized in Figure 5. However, technical metadata might have a much stronger impact when analyzing raw images or rendered images instead.

We finally show in Section 6 that the semantically refined skin color models improve skin detection in images. Such skin maps are important for subsequent image processing tasks such as face recognition, hand detection, or color correction.

In future work we plan to investigate multiple metadata attributes together instead of one at a time. This is important as images usually come with diverse metadata, and color models can be better refined if they are all incorporated into the model.

ACKNOWLEDGMENTS

This study is supported by the research grant for ADSC's Human Sixth Sense Programme from Singapore's Agency for Science, Technology and Research (A*STAR).

REFERENCES

[1] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K., "Attribute and Smile Classifiers for Face Verification," in [*IEEE International Conference on Computer Vision*], (2009).

[2] "Google custom search API." https://developers.google.com/custom-search/docs/xml_results (last checked, Nov 2013).

[3] Nickerson, D., "History of the Munsell color system and its scientific application," *Journal of the Optical Society of America* **30**(12), 575–586 (1940).

[4] Bartleson, C. J., "Memory colors of familiar objects," *Journal of the Optical Society of America* **50**(1), 73–77 (1960).

[5] Yendrikhovskij, S. N., Blommaert, F. J. J., and de Ridder, H., "Color reproduction and the naturalness constraint," *Color Research & Application* **24**(1), 52–67 (1999).

[6] Park, D., Kwak, Y., Ok, H., and Kim, C. Y., "Preferred skin color reproduction on the display," *Journal of Electronic Imaging* **15**(4) (2006).

[7] Zhang, X., Jiang, J., Liang, Z., and Liu, C., "Skin color enhancement based on favorite skin color in HSV color space," *IEEE Transactions on Consumer Electronics* **56**(3), 1789–1793 (2010).

[8] You, J. and Chien, S., "Saturation enhancement of blue sky for increasing preference of scenery images," *IEEE Transactions on Consumer Electronics* **54**(2), 762–768 (2008).

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

[10] Yogarajah, P., Condell, J., Curran, K., Cheddad, A., and McKevitt, P., "A dynamic threshold approach for skin segmentation in color images," in [*IEEE International Conference on Image Processing*], 2225–2228 (2010).

[11] Lakshmi, H. C. V. and PatilKulakarni, S., "Segmentation algorithm for multiple face detection for color images with skin tone regions," in [*International Conference on Signal Acquisition and Processing*], 162–166 (2010).

[12] Erdem, C. E., Ulukaya, S., Karaali, A., and Erdem, A. T., "Combining Haar feature and skin color based classifiers for face detection," in [*IEEE International Conference on Acoustics, Speech and Signal Processing*], 1497–1500 (2011).

[13] Xie, S. and Pan, J., "Hand detection using robust color correction and gaussian mixture model," in [*International Conference on Image and Graphics*], 553–557 (2011).

[14] Liu, L., Sang, N., Yang, S., and Huang, R., "Real-time skin color detection under rapidly changing illumination conditions," in [*IEEE Transactions on Consumer Electronics*], 1295–1302 (2011).

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

[16] de Rigal, J., des Mazis, I., Diridollou, S., Querleux, B., Yang, G., Leroy, F., and Barbosa, V. H., "The effect of age on skin color and color heterogeneity in four ethnic groups," *Skin Research and Technology* **16**(2), 168–178 (2010).

[17] Huiskes, M. J., Thomee, B., and Lew, M. S., "New trends and ideas in visual concept detection," in [*ACM International Conference on Multimedia Information Retrieval*], (2010).

[18] Viola, P. and Jones, M. J., "Robust real-time face detection," *International Journal of Computer Vision* **57**(2), 137–154 (2004).

[19] "OpenCV." http://opencv.org (last checked, Nov 2013).

[20] "Flickr API." http://www.flickr.com/services/api/ (last checked, Nov 2013).

[21] Huiskes, M. J. and Lew, M. S., "The MIR flickr retrieval evaluation," in [*ACM International Conference on Multimedia Information Retrieval*], (2008).

[22] Wilcoxon, F., "Individual comparisons by ranking methods," *Biometrics Bulletin* **1**(6), 80–83 (1954).

[23] Lindner, A., *Semantic Awareness for Automatic Image Interpretation*, PhD thesis, EPFL (2013).