A Comparative Evaluation of Three Skin Color Detection Approaches

Dennis Jensch*, Daniel Mohr** , Gabriel Zachmann**

* Clausthal University of Technology
email: dennis.jensch@tu-clausthal.de
** University of Bremen
email: {mohr,zach}@informatik.uni-bremen.de

Abstract

Skin segmentation is a challenging task due to several influences such as unknown lighting conditions, skin colored background, and camera limitations. A lot of skin segmentation approaches were proposed in the past including adaptive (in the sense of updating the skin color online) and non-adaptive approaches.

In this paper, we compare three skin segmentation approaches that are promising to work well for hand tracking, which is our main motivation for this work. Hand tracking can widely be used in VR/AR e.g. navigation and object manipulation.

The first skin segmentation approach is a well-known non-adaptive approach. It is based on a simple, pre-computed skin color distribution. Methods two and three adaptively estimate the skin color in each frame utilizing clustering algorithms. The second approach uses a hierarchical clustering for a simultaneous image and color space segmentation, while the third approach is a pure color space clustering, but with a more sophisticated clustering approach.

For evaluation, we compared the segmentation results of the approaches against a ground truth dataset. To obtain the ground truth dataset, we labeled about 500 images captured under various conditions.

First presented at the Workshop Virtuelle und Erweiterte Realität, 2012, extended and revised for JVRB

1 Introduction

Skin color is an important and powerful feature for several applications, e.g. image classification (decide if human beings are found in the image or not), face detection, and hand pose estimation. To detect the skin color in images, a lot of approaches have been proposed.[1] Basically, all approaches try to learn the skin color, and then use the distribution to classify the images. Most approaches perform the classification independently to each pixel because of its low computation time. The most challenging task is to learn the skin color distribution due to its dependence on a lot of parameters like ethnicity, illumination conditions, and camera limitation (limited dynamic range, color distortion, and so forth). Furthermore, skin color is often also found in the background, which yields a lot of clutter in the segmentation result.

Skin segmentation approaches can be classified into adaptive and non-adaptive methods. Non-adaptive methods learn the skin color distribution offline. During tracking, this distribution is used for classification. Such approaches have the main problem that they have a low segmentation quality under varying conditions. Adaptive methods try to overcome this problem by updating the distribution online.

The segmentation quality is crucial for example for hand and face tracking approaches which often use the segmentation foreground as silhouette. Thus, the better the segmentation results are, the faster and more reliable the tracking will work. For this reason, we want to evaluate and compare three different skin segmentation approaches which we expect to work well

[1] The most often used term for this task are *skin detection* and *skin segmentation.*

urn:nbn:de:0009-6-348, ISSN 1860-2037

for tracking purposes.

Our first candidate is a well-known skin segmentation approach proposed by Rehg and Jones [JR02]. This approach uses skin and background color distributions to classify the image pixel-wise as skin and non-skin. The color distributions are learned offline. Consequently, the approach is non-adaptive and does not perform well under non-static conditions, e.g. illumination changes.

Our second candidate [MZ07] formulates the segmentation as a combined image and color space clustering. Basically, the result is an image segmentation. Each image region is then classified as skin or non-skin by utilizing a precomputed training vector.

Our third candidate is the approach proposed by [MZ07]. The approach uses combined image and color space clustering, which we have slightly optimized to achieve better segmentation results. We will explain the modifications in detail in Section 3.

For evaluation, we compared the approaches against each other using eleven different *skin thresholds* to reveal the relations between the false positives and false negatives generated by all approaches. *False positives* are background pixels that are falsely classified as foreground, and *false negatives* are skin pixels that are falsely classified as background. The segmentation results are visually plotted as color coded images. We also count the number of correctly and incorrectly classified skin pixels and discuss the results using receiver operating characteristic (ROC) curves. Next, we interpret the curves, analyze certain characteristics of the approaches, and compare their results.

In Section 3, we will first give a short overview of the approaches we want to evaluate and compare against each other, and in Section 4 we will describe our ground truth dataset and the test setup.

2 Related Work

A lot of segmentation approaches have been proposed. Two kind of methods are graph cutting and color space clustering. Graph cutting methods use the images as weighted graphs and map segmentation to graph cutting with specific cost functions. The color space clustering methods segment images using models in color space.

The approach of Rital and Miguet [RCM05] used a multilevel hypergraph. They segmented gray level images with regards to its performance on noisy images.

As examples for color space clustering approaches, Zhang [ZW00] presented a k-mean clustering in the HSI (hue, saturation and intensity) color space for application to medical images.

Kim and Lee [KLL04] introduced an initialization scheme for fuzzy c-means clustering using reference colors to calculate dominant color; defined as most distinguishable colors. The initial centroids were selected from the colors closest to the dominant colors.

Rehg and Jones [JR02] proposed a skin detection approach that utilizes a histogram-based and mixture model representation of skin and non-skin color. Color models for skin and non-skin classes were constructed from a dataset of one billion hand labeled pixels. With large training data this histogram based representation is superior while with small training data the mixture model provide better segmentation results. The approach provides a detection rate of 80 % for web images. The inflexibility of a static color model is a disadvantage that could yield less performance on images with different conditions than their training data set.

To obtain new estimations of the color distribution during run-time, Sigal et al. [SSA00] updated the skin color using the segmentation result of the previous frame after some post-processing (e.g. morphological filter for noise reduction, hole filling, and erasing small isolated skin colored regions). The drawback of this approach is that the skin color estimate converges to background color if the initial estimation is not of high quality.

[DGV04] improved skin detection by a variational EM algorithm with spatial constraints. For initialization, they used the skin color model of [JR02].

The Berkeley segmentation dataset and benchmark [MFTM01] provides a large set of ground truth data. It can be used for different segmentation approaches, resulting in a growing database of comparable benchmark results. For the benchmark, results of different thresholds are gathered and used to generate a precision-recall curve.

A survey by Vezhnevets et al. [VSA03] examined several pixel-based skin detection approaches. They analyzed and compared their characteristics and provide an overview of all approaches and how well they are suited for several applications.

A survey by Kakumanu et al. [KMB07] presented, evaluated and compared various candidates for three aspects concerning skin color detection. These aspects were the chosen color space (for example RGB, HSI and YUV), skin color classifier (for example

histogram-based approaches, gaussian mixture models and artificial neural networks) and possible illumination adaption approaches (for example gray world approaches) which reduce problems with bad illumination conditions. The survey serves as a good overview for developers of new skin detection approaches, presenting a wide variety of methods that could be used.

In contrast to previous work, our goal is to evaluate and compare skin detection approaches for application to human hand tracking under various conditions e.g. skin colored background, bad illumination, and over- and underexposured skin. For this reason, we generate new ground truth datasets that fulfill these conditions.

3 Skin Segmentation Approaches

The first skin segmentation approach uses a static color model. In contrast, the other two approaches we evaluate, learn the skin color online. The third method is an adaption of the second one and therefore they are rather similar. One of their differences is the way to determine the number of clusters the image is partitioned into. While the second method uses a hierarchical subdivision to determine the number of cluster that yield the best results, the third method tests several number of clusters and selects the best option based on some criteria that are explained in detail below (Sec. 3.3).

3.1 RehgJones

The first approach was proposed by Rehg and Jones [JR02], which we will denote in this paper as *RehgJones*. They used a huge hand-segmented image database from the internet. With this database a color distribution was trained and represented as 3D histograms. The approach then used the histograms to classify pixels as skin or non-skin. First, they used a 3D color histograms with 256^3 bins. They tested different numbers of bins for the color distribution representation. It turned out that 32 bins per channel yield the best segmentation result.

They also tested a Gaussian mixture model for skin color representation. The segmentation quality was lower compared to the histogram based color representation.

All color distributions (histograms and GMM) are computed in the RGB color space which is in contrast to many other approaches that convert the image into perception oriented color spaces.

Based on this histograms the likelihood of a color rgb to belong to skin or background can be computed by

$$P(rgb|skin) = \frac{s[rgb]}{T_s} \tag{1}$$

$$P(rgb|\neg skin) = \frac{n[rgb]}{T_n} \tag{2}$$

where $s[rgb]$ indicates the pixel count contained in the bin rgb within the skin histogram, and $n[rgb]$ is the respective equivalent in the background histogram.

The normalization factors T_s and T_n contain the total number of color pixel counts in the skin- and background histograms.

The final likelihood L_{skin} of a color pixel to be skin can be computed by the ratio of a pixel to be found in skin regions and background regions:

$$L_{skin}(rgb) = \frac{P(skin|rgb)}{P(\neg skin|rgb)} \tag{3}$$

$$= \frac{P(rgb|skin)P(skin)}{P(rgb|\neg skin)P(\neg skin)} \tag{4}$$

[JR02] proposed to choose

$$P(skin) = \frac{T_s}{T_s + T_n} \tag{5}$$

as reasonable choice of priors and, of course, $P(\neg skin) = 1 - P(skin)$. If a binary classification is necessary, a color value is classified as skin if

$$L_{skin}(rgb) > \tau \tag{6}$$

In our experiments, we use a value θ (defined in Sec. 4.2) similar to τ, but we ensure that θ is normalized to $[0, 1]$ for the sake of comparability to the other approaches.

3.2 HybridClustering

The second approach was developed by Mohr and Zachmann [MZ07], which we will denote as *HybridClustering*. The approach formulates the segmentation as a combined color and image space clustering. To compute the clusters, the expectation maximization (EM) algorithm [Bil98] is utilized. In each EM step, first, the clustering is applied to the pixels in

(a) Input image (b) Result of a clustering (c) Edge image

Figure 1: We use three different quality measures to determine the optimal number of clusters an image (a) has to be split into. The first two quality measures are based on the re-projection of the clustering result to image space. Image (b) shows an example: each individual color represents a cluster id. While the first measure uses the border length itself between the individual clusters, the second approach measures the average edge response (c) at the cluster borders. The third quality measure tests the closeness of the color pixels to the cluster centers in color space.

color space, and then the pixel-to-cluster correspondence modified by an image space neighborhood condition. Each cluster represents a homogeneous colored image region. The regions are classified as skin and non-skin based on a precomputed training vector. To reduce the number of isolated pixels or very small regions in image space, the approach uses a spatial constraint to modify the corresponding pixel probabilities during the clustering process appropriately. First, edges are extracted using the Laplace edge filter, and second, an edge distance map is computed using the inverse distance weighted edge intensities in a local neighborhood.

The idea behind the smoothing is that pixels in a local neighborhood without edges belong to the same object in the image, and thus, should belong to the same image cluster (in color space). Thus, in such regions, the pixel-to-cluster probabilities are smoothed. This leads to the following smoothing operation

$$p_n(\mathbf{x}_i|\Sigma_j) = p(\mathbf{x}_i|\Sigma_j)D(\mathbf{x}_i) + (1 - D(\mathbf{x}_i))\bar{p}(\mathbf{x}_i|\Sigma_j) \quad (7)$$

where $D(\mathbf{x}_i)$ is the edge distance image, Σ_j are the parameters (mean and covariance matrix) of cluster j, $p(x_i|\Sigma_j)$ is the probability that $\mathbf{x}_i$ belongs to cluster j and $\bar{p}(\mathbf{x}_i|\Sigma_j)$ is the average probability that all pixels in the neighborhood of $\mathbf{x}_i$ belong to cluster j.

To determine the "best" number of clusters, they have chosen to use a divisive clustering approach because divisive clustering allows to early skip clusters with too low a skin probability, which can significantly reduce the computation time. By contrast, agglomerative clustering approaches would not allow to save computation time as described above. The stopping criterion for a further subdivision during the divisive clustering is based on the edge distance map. The fewer edges on the cluster borders in image space are found, the lower the probability that the subdivision guided by the clustering cuts two distinct image objects.

3.3 NeuralGasColorClustering

The third approach is an adaptation of [MZ07], which we will denote as *NeuralGasColorClustering*. We have developed it to reveal the influence of the clustering approach to the final segmentation quality. For this purpose, we replaced the EM algorithm by the matrix neural gas (MNG) method [AH10]. The main advantage of MNG is that it is much more robust with respect to initialization. Thus, the approach more often converges to the global maximum compared to the EM algorithm. To keep this nice property, we have not applied image space smoothing as is done in [MZ07]. Furthermore, we have replaced the way to determine the number of clusters that performs best: whereas [MZ07] uses a hierarchical subdivision, we tested several numbers of clusters and chose the best one, i.e. first, we cluster the image into $k = 2$ clusters, then we evaluate the quality of the result, and then use $k = 3$ clusters and so forth up to a limit of n clusters. In or-

Tag	Description
Set_A, Set_I, Set_N	complex BG, bad illumination
Set_B, Set_C, Set_G, Set_H, Set_J, Set_K, Set_L	simple BG, good illumination
Set_D, Set_E, Set_F	simple BG, bad illumination
Set_M, Set_O	complex BG, good illumination

Table 1: An overview of our ground truth datasets. We captured image sequences under different illumination conditions and with simple and complex background including skin colored background.

Algorithm 1: BorderLength(C)

Input: cluster image C

Output: cluster quality measure Q

borderLength = 0

foreach $\mathbf{x} \in C$ **do**

  if $\exists \mathbf{y} \in \mathcal{N}(\mathbf{x})$: $C(\mathbf{x}) \neq C(\mathbf{y})$ **then**

    borderLength+=1

Q = 1 - $\frac{\text{borderLength}}{\text{size}(C)}$

der to determine the best number of clusters from this results, we need a measure to compute the quality of the clustering result. We tested three different quality measures.

We use the following notations to explain the quality measures:

- I denotes the original input image. An example is shown in Fig. 1a.
- E denotes the edge intensity image of the original image. An example is shown in Fig. 1c.
- μ_i and Σ_i are the cluster mean value and covariance matrix for cluster i, obtained by applying MNG to the input image I
- C is a mapping function from an image pixel to the corresponding cluster index. An example image is shown in Fig. 1b. The clusters indices are color coded.
- $\mathcal{N}(\mathbf{x})$ denotes the set of pixels in the direct neighborhood of $\mathbf{x}$.

The first quality measure, *Border Length* (BL), measures the length of the obtained cluster borders in image space (image region). The shorter the borders are, the more compact the image region we expect to be, and consequently the better the clustering result is. We compute the measure as shown in Algorithm 1.

The second quality measure, *Border Edges* (BE) does not use the border length itself but the edge response (obtained by an edge detector) across the borders. The idea of this measure is a good clustering should separate objects in an image, and at such borders between object we usually observe the strongest edge response. Higher values denote a better clustering quality. The pseudo code is shown in Algorithm 2.

Algorithm 2: BorderEdges(C,E)

Input: cluster image C, edge image E

Output: cluster quality measure Q

borderLength = 0

edgeResponse = 0

foreach $\mathbf{x} \in C$ **do**

  if $\exists \mathbf{y} \in \mathcal{N}(\mathbf{x})$: $C(\mathbf{x}) \neq C(\mathbf{y})$ **then**

    borderLength+= 1

    edgeResponse += E($\mathbf{x}$)

Q = $\frac{\text{edgeResponse}}{\text{borderLength}}$

The third quality measure, *Color Space Compactness* (CSC) operates in color space, and tests the proximity of all pixels to the corresponding cluster center using the Mahalanobis distance. The matrix for the Mahalanobis distance is computed by the MNG algorithm. Please see Algorithm 3 for the detailed description of the measure.

The three measures, of course, can also be combined into a single measure, e.g. by a weighted sum of the individual measures.

4 Test Setup

In this section, we will first describe our ground truth dataset, and then, explain the methods we use for evaluation.

Algorithm 3: ColorSpaceCompactness(I,C,μ_i, Σ_i)

Input: original Image I, cluster image C, cluster parameters μ_i, Σ_i

Output: cluster quality measure Q

sumProb = 0

foreach $\mathbf{x} \in I$ **do**

  avgDist = 0

  foreach *clusters* i **do**

    avgDist += $(\mathbf{x} - \mu_i)^{\mathsf{T}} \Sigma_i^{-1} (\mathbf{x} - \mu_i)$

  avgDist /= #clusters

  dist = $(\mathbf{x} - \mu_{C(\mathbf{x})})^{\mathsf{T}} \Sigma_i^{-1} (\mathbf{x} - \mu_{C(\mathbf{x})})$

  sumProb += $1 - \frac{dist}{avgDist}$

Q = $\frac{sumProb}{size(I)}$

4.1 Ground Truth Data

To be able to evaluate and compare skin detection approaches, we need a ground truth dataset with varying skin color and background conditions, such that the ground truth dataset is representative for real applications. We have decided to generate our own ground truth dataset because our focus is hand tracking and ,thus, we need a skin segmentation approach that works best for typical video sequences showing a hand in the image. Segmentation approaches that are optimized to work well for other images/videos are of low interest for us and could even yield for lower segmentation quality for images showing a hand.

To obtain the ground truth dataset, we manually labeled a large number of images. The ground truth dataset consists of 15 subsets. For simplicity, we denote each of this subsets as "dataset" and write explicitly "ground truth dataset" if we refer to the whole ground truth dataset. All datasets consist of images showing a single person at different postures and under different background and illumination conditions. The ground truth dataset consists of 483 images. Five datasets contain a *complex* background. With complex background we mean that several objects are visible in the background, potentially skin colored or highly textured. In contrast, the other datasets have a *simple* background. *Simple* means that the whole background has a homogeneous color. Six datasets have bad illumination conditions. A detailed overview of the conditions for all datasets is shown in Table 1, and example pictures for each dataset are given in Fig. 9. We expect that the datasets with a complex background and/or bad illumination conditions are more challenging for the segmentation approaches. The segmentation quality of these datasets are of special interest in our work.

4.2 Evaluation Methods

In this paper, we use the following notations:

- *False positives* are background pixels that are classified as skin,
- *false negatives* are skin pixels that are classified as background, and
- *true positives* and *true negatives* are correctly classified pixels.

Fig. 2 illustrates the four pixel types by an example.

Please note, that the skin segmentation approaches compute for all image pixels a probability to be skin color. In order to be able to compute *false positives*, *false negatives* etc., we have to binarize the probabilities i.e. convert the skin probabilities to binary values. The threshold used for binarization basically controls the trade off between the false negatively and false positively classified pixels. In the following we denote this threshold simply as *skin threshold* θ.

For evaluation, we use receiver operating characteristic (ROC) curves. ROC curves visualize the relationship between false positives and true positives. Different relations between false and true positives are

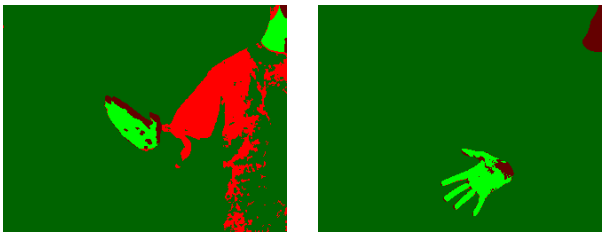
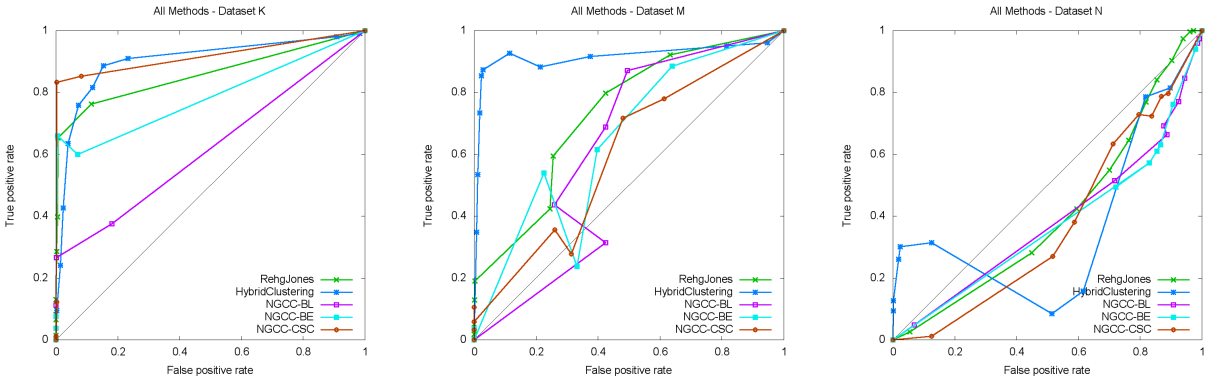


Figure 2: Segmentation results as color-coded images. The left image shows a detection with mixed quality. Skin is mostly detected (true positive; light green), but also large regions of non-skin is classified as skin (false positive; light red); The right image shows a detection with nearly perfect classified non-skin (true negative; dark green). Some skin regions are falsely classified as non-skin (false negative; dark red).



(a) ROC curves for dataset K (b) ROC curves for dataset M (c) ROC curves for dataset N

Figure 4: Comparing the ROC curves for three datasets with different backgrounds (simple background, cluttered background and skin colored background) shows that none of the skin detection approaches is superior. The approach performing best depends on the individual dataset and the false positive and true positive rate as well.

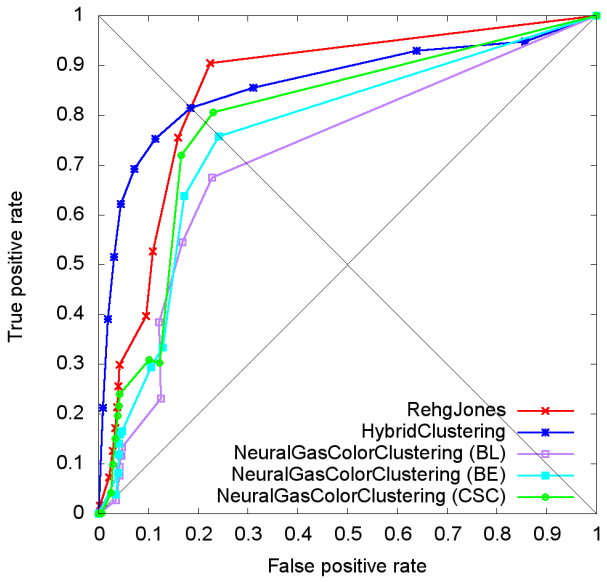


Figure 3: We evaluate the segmentations approaches by operating characteristic (ROC) curves analysis. ROC curves visualize the relationship between false positives and true positives. The closer the curve is to the y-axis on the left, the better the approach is.

generated by updating the skin threshold θ. We evaluated each approach using 11 different values for θ from 0.01 up to 0.9. For the approach *NeuralGasColorClustering* we set the cluster parameter k to a maximum of eight.

The trivial result for $\theta = 1$ is located in the origin i.e. no pixels are classified as skin. Of course, no evaluation has to be done for this point. The same holds for $\theta = 0$. In this case, all pixels are classified as skin (found at coordinate $(1, 1)$ in the ROC curves). Obviously, the results for the 11 runs are located between this two points. Higher thresholds are closer to the origin.

Additionally we compute the Equal Error Rate (EER). The EER is the segmentation quality with equal false positive and false negative rate. It can easily be read out from the ROC curves by intersecting the curve of interest with the secondary diagonal.

5 Results

In this section, we will first discuss the quality of all three approaches using ROC curve analysis. Then we will further investigate the *NeuralGasColorClustering* approach proposed in this paper. Particularly, we are interested in the number of clusters the approach divides each image into. We want to compare the behavior of the three cluster quality measures and a potential over- or underestimation of the best number of clusters. We can utilize this in the future to adapt the

minimum and maximum allowed number of clusters. Finally, we measure the computation time for all approaches which is of high interest for real-time applications.

5.1 Segmentation Quality

In Fig. 3, we observe that the *HybridClustering* approach performs best on average because the ratio between the true positives and false positives is higher compared to the other approaches except for the lowest three values for θ. But in real applications we do not want such a high false positive rate. Surprisingly, *RehgJones* is superior compared to *NeuralGasColorClustering*.

Comparing the ROC curves of *NeuralGasColorClustering* using the three different methods (BL, BE and CSC) to determine the "best" number of clusters, we observed that CSC yields the best ratio between true positives and false positives. We have also tested a linear combination of all three cluster quality measures, but we observed no increase in quality.

Both, *RehgJones* and *HybridClustering*, have the same equal error rate of about 82 %. The equal error rate for the *NeuralGasColorClustering* are slightly lower and depend on the cluster subdivision criteria which are 70 % for BL, 75 % for BE and 78 % for CSC.

So far, we have discussed the overall quality of the skin segmentation approaches using the whole ground truth dataset. Next, we want to analyze the segmentation quality for the individual datasets. This is important because we have datasets with different illumination conditions and background.

First, we observed that the best skin detection approach varies from dataset to dataset as one can see in Fig. 4. Thus, none of the skin detection approaches is superior. For example, in Fig. 4a, *NeuralGasColorClustering* performs better for a lower false positive and true positive rate, while for higher false positive and true positive rates, *HybridClustering* performs better. In Fig. 4b, *HybridClustering* is superior, and in Fig. 4c there is no clear winner at all.

Second, we have observed a high variation between the ROC curves of the individual datasets. For the discussion of the results of the individual datasets, we have decided to pick three representative sets for different classes of datasets, homogeneous background (Set_K), textured background (Set_M) and skin colored background (Set_N) because we would get no significant additional information if discussing each of the 15 datasets separately. Additionally, plotting all 15 datasets would result in extremely unclear figures.

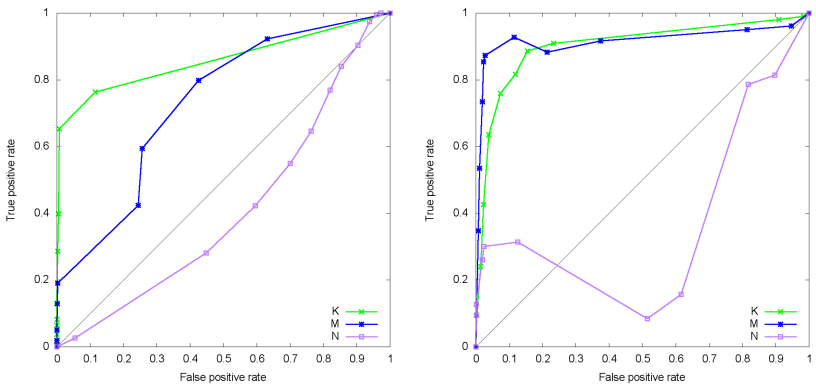
For *RehgJones*, we observed a moderate variance between individual datasets (Fig. 5a) Generally, datasets with homogeneous background have a better ration between true positives and false positives, and complex background yield a worse ratio (for example images for such datasets see Fig. 9a, Fig. 9i and Fig. 9o).

HybridClustering has the lowest variance between individual datasets (Fig. 5b). The big exception is Set_N, which we will discuss in detail below. In summary, the approach has a high detection rate for all datasets. Not surprisingly, the best results are achieved with simple background and normal illumination (for example Set_K).

For *NeuralGasColorClustering*, we observed the highest variance (Fig. 6). The segmentation result of *NeuralGasColorClustering* strongly depends on the matrix neural gas clustering. MNG randomly initializes the prototype positions. Despite the better convergence behavior of MNG, the prototype positions still vary from run to run and lead to slightly different final clusters and consequently different segmentation results. Small variation in the segmentation result can lead the individual image regions to be classified differently (i.e. as skin at one run, and as background at the next run) if their average color value is close to another color that has a different classification. We could expect the same behavior in the *HybridClustering* approach, but think that the spatial smoothing alleviates this "alternating" effect. We have actually observed a strong variation of the segmentation results. Because performing multiple runs for all approaches, for all datasets, and all values for θ is by far too time consuming, we have decided to pick Set_M to visualize the amount of variation. We performed four runs with identical input values.

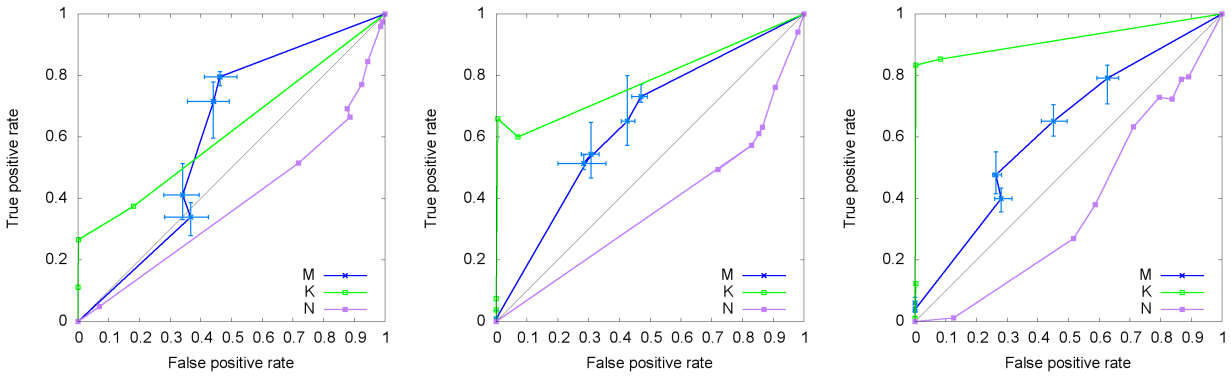
It is no big surprise that datasets with complex backgrounds have the highest variance. In these datasets, different prototype positions lead more likely to different clusters because the complex background has more chances for different partitionings. *NeuralGasColorClustering* provides good results for the datasets with mainly simple background and good illumination conditions except Set_N.

Set_N yields by far the worst segmentation quality. The main reason is the red skin color-like door in the



(a) RehgJones (b) HybridClustering

Figure 5: We have chosen three individual datasets to analyze the influence of different illumination and background conditions to the segmentation quality. The ROC curves show that the variation from the overall ROC curve in Fig. 3 is moderate except Set_N, which contains a large skin colored background region.



(a) Border Length (b) Border Edges (c) Color Space Compactness

Figure 6: We have chosen three individual datasets to analyze the influence of different illumination and background conditions to the segmentation quality. The above images show the results for all three measures to determine the best number of clusters in *NeuralGasColorClustering*. We observed a high variance between the individual datasets. Even for multiple runs of the same dataset with the identical input parameters, we observe a high variance, in particular for Set_N, which is visualized with additional error bars.

(a) Original image (b) Result for $\theta = 0.30$ (c) Result for $\theta = 0.50$

Figure 7: *HybridClustering* yields better quality for Set_N than the other approaches. But this is limited to higher values of the skin classification threshold θ (defined in Sec. 4.2). In (b), most of the skin regions are not classified correctly, instead the background is identified as skin. The other both approaches achieve results similar to (b). Image (c) shows a significantly better quality for this example achieved using *HybridClustering*.

background, which is classified as skin by all three approaches. Additionally, the hand is overexposed due to the bad illumination conditions, and thus, several hand pixels are white, which is also hard to be distinguished from the white shirt. For this dataset, *HybridClustering* achieved better results than the other methods (Fig. 7).

5.2 Optimal Number of Clusters for *NeuralGasColorClustering*

We have also recorded the number of clusters actually chosen by the three cluster quality measures because they are crucial for the overall segmentation quality. Figure 8 shows the number of clusters computed by the individual measures for all images in our ground truth dataset.

Border Length considers only the border length and selects the lowest number of clusters compared to the other measures. CSC always yields the highest number of clusters with values most often between 6 and 8. Interestingly, *NeuralGasColorClustering* with CSC always has a better segmentation quality than *NeuralGasColorClustering* with BL or BE. Thus, we suppose that on average six to eight clusters perform best for application to skin detection.

The three quality measures can be easily combined by using the weighted sum of the three measures. Determining the best weights, of course, is not trivial. We have tested a combination with equal weights, but we observed no increase in quality. We also do not expect to get a significant improvement using other weights, because *Color Space Compactness* is superior compared to the other two weights.

5.3 Computation Time

For each approach we measured and averaged the computation time of three runs. The results are shown in Table 2. The clustering-based approaches have a significantly higher computation time because the clustering itself is very expensive. We observed the highest computation time for *NeuralGasColorClustering*. The main reason is that for a selected value *n* of the parameter k (see Section 3) the approach has to perform the image clustering *n* times.

We want to mention that we used an unoptimized C++ implementation of all approaches. A high amount of speedup could be achieved by parallelizing the algorithms according to the massively parallel programming paradigm.

Table 2 shows that the influence of the cluster quality measure on the computation time of *NeuralGasColorClustering* is less than 2 %.

In contrast, the hierarchical clustering approach in *HybridClustering* allows the approach to early prune large parts of the image pixels, which keeps the computation time low.

Clearly, *RehgJones* has the lowest computation time because only a histogram lookup has to be done for segmentation. But, of course, it is not fair to compare *RehgJones* with the other approaches because, in contrast to the other approaches, the skin color distribution of *RehgJones* is not able to adapt to varying con-

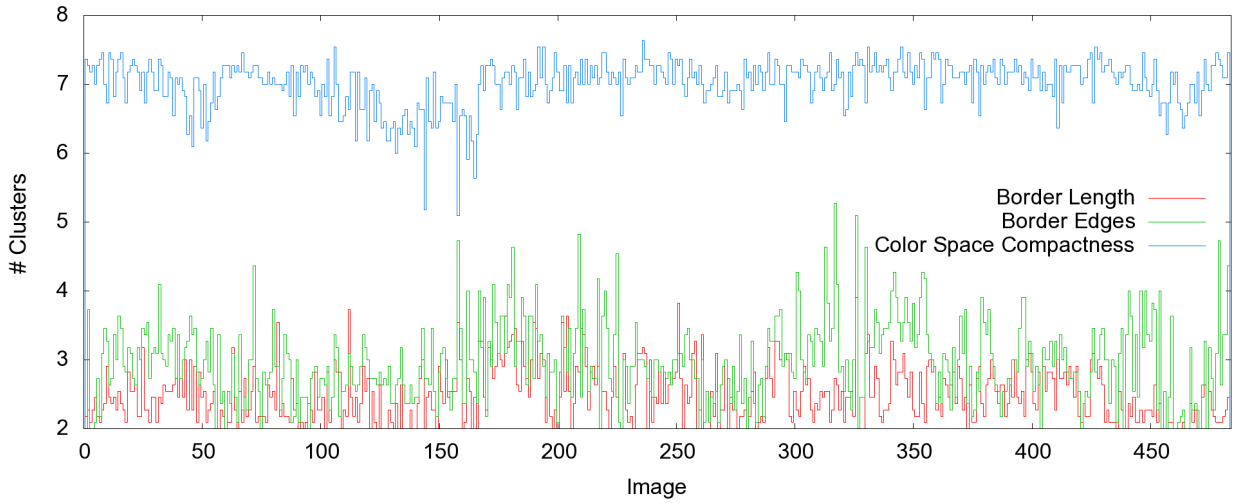


Figure 8: For *NeuralGasColorClustering*, we used three different methods (*Border Length* (BL), *Border Edges* (BE) and *Color Space Compactness* (CSC)) to determine the optimal number of clusters. The plot shows for the three methods and all images in the ground truth dataset the number of clusters chosen. Interestingly, CSC always decides a much higher number of clusters to be best in contrast to BL and BE.

Approach	Time (ms)	Std. dev. (ms)
RehgJones	1.23	0.06
HybridClustering	508.79	442.39
$NeuralGasColorClustering_{BL}$	45 013.95	2 458.10
$NeuralGasColorClustering_{BE}$	45 886.43	2 635.47
$NeuralGasColorClustering_{CSC}$	45 460.82	2 961.72

Table 2: Computation time for each segmentation approach, averaged over 3 runs.

ditions because of the color distribution that is learned in a pre-processing step and not adapted to any input image. The other both approaches can adapt to some amount because of the clustering step keeps skin and non-skin regions together and then classifies this regions as a whole.

6 Conclusion

We compared the quality of three skin segmentation approaches, *RehgJones*, *HybridClustering*, and *NeuralGasColorClustering*, by way of ROC curves. Additionally, we measured the computation time to evaluate their usefulness for real applications.

We observed that all three approaches provide a good quality for datasets with simple background and a lower quality for datasets with a complex background. The *NeuralGasColorClustering* also has some difficulties with complex backgrounds. *RehgJones* and *HybridClustering* provide the highest true positive rate but also a high false positive rate. On average, *HybridClustering* performed best and *NeuralGasColorClustering* worst.

On *NeuralGasColorClustering* we observed a lower true positive and false positive rate. For the *NeuralGasColorClustering* approach, *Color Space Compactness*, which determines the number of clusters, has turned out to be superior.

In the future, we plan to investigate whether the low false positive rate of *NeuralGasColorClustering* could be advantageous for motion tracking despite of its low true positive rate. Additionally, we want to further investigate whether a higher number of clusters yield better segmentation results. For this purpose, we need to perform more tests with more complex images and higher number of clusters.

References

[AH10] Banchar Arnonkijpanich and Barbara Hammer, *Global coordination based on matrix neural gas for dynamic texture synthesis*, Artificial Neural Networks in Pattern Recognition - 4th IAPR TC3 Workshop, ANNPR 2010, Cairo, Egypt, April 11-13, 2010. Proceedings (Berlin, Heidelberg), Lecture Notes in Computer Science Volume 5998, Springer-Verlag, 2010, DOI 10.1007/978-3-642-12159-3_8, pp. 84–95, ISBN 978-3-642-12158-6.

[Bil98] Jeff Bilmes, *A Gentle Tutorial of the EM algorithm and its application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, Tech. Report TR-97-021, International Computer Science Institute ICSI, 1998.

[DGV04] A. Diplaros, T. Gevers, and N. Vlassis, *Skin detection using the EM algorithm with spatial constraints*, IEEE International Conference on Systems, Man and Cybernetics, vol. 4, 2004, DOI 10.1109/ICSMC.2004.1400810, pp. 3071–3075, ISBN 0-7803-8566-7.

[JR02] Michael J. Jones and James M. Rehg, *Statistical Color Models with Application to Skin Detection*, International Journal of Computer Vision **46** (2002), no. 1, 81–96, ISSN 0920-5691, DOI 10.1023/A:1013200319198.

[KLL04] Dae-Won Kim, Kwang Hyung Lee, and Doheon Lee, *A novel initialization scheme for the fuzzy c-means algorithm for color clustering*, Pattern Recognition Letters **25** (2004), no. 2, 227–237, ISSN 0167-8655, DOI 10.1016/j.patrec.2003.10.004.

[KMB07] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, *A survey of skin-color modeling and detection methods*, Pattern Recognition **40** (2007), no. 3, 1106–1122, ISSN 0031-3203, DOI 10.1016/j.patcog.2006.06.010.

[MFTM01] D. Martin, C. Fowlkes, D. Tal, and J. Malik, *A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics*, Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings., vol. 2, 2001, DOI 10.1109/ICCV.2001.937655, pp. 416–423, ISBN 0-7695-1143-0.

(a) Set_A (b) Set_B (c) Set_C

(d) Set_D (e) Set_E (f) Set_F

(g) Set_G (h) Set_H (i) Set_I

(j) Set_J (k) Set_K (l) Set_L

(m) Set_M (n) Set_N (o) Set_O

Figure 9: The main differences between the image sequences in our ground truth database are the background (complex or simple) and illumination conditions, which results into four image categories. The above images show examples for the datasets Set_A to Set_O. Each set consists of 10–20 images.

[MZ07] Daniel Mohr and Gabriel Zachmann, *Segmentation of Distinct Homogeneous Color Regions in Images*, The 12th International Conference on Computer Analysis of Images and Patterns (CAIP) (Vienna, Austria), Lecture Notes in Computer Science Volume 4673, Springer Verlag, 2007, DOI 10.1007/978-3-540-74272-2_54, pp. 432–440, ISBN 978-3-540-74271-5.

[RCM05] Soufiane Rital, Hocine Cherifi, and Serge Miguet, *A segmentation algorithm for noisy images*, 205–212, DOI 10.1007/11556121_26.

[SSA00] Leonid Sigal, Stan Sclaroff, and Vassilis Athitsos, *Estimation and Prediction of Evolving Color Distributions for Skin Segmentation Under Varying Illumination*, IEEE Conference on Computer Vision and Pattern Recognition, 2000. Proceedings, vol. 2, 2000, DOI 10.1109/CVPR.2000.854764, pp. 152–159, ISBN 0-7695-0662-3.

[VSA03] Vladimir Vezhnevets, Vassili Sazonov, and Alla Andreeva, *A Survey on Pixel-Based Skin Color Detection Techniques*, GraphiCon, 2003, pp. 85–92.

[ZW00] C. Zhang and P. Wang, *A New Method of Color Image Segmentation Based on Intensityand Hue Clustering*, 15th International Conference on Pattern Recognition, 2000. Proceedings., vol. 3, 2000, DOI 10.1109/ICPR.2000.903620, pp. 613–616, ISBN 0-7695-0750-6.

Citation

Dennis Jensch, Daniel Mohr, and Gabriel Zachmann, *A Comparative Evaluation of Three Skin Color Detection Approaches*, Journal of Virtual Reality and Broadcasting, 12(2015), no. 1, January 2015, urn:nbn:de:0009-6-40888, ISSN 1860-2037.