A survey on visual adult image recognition

Christian X. Ries, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Ries, Christian X., and Rainer Lienhart. 2012. "A survey on visual adult image recognition." *Multimedia Tools and Applications* 69 (3): 661–88.
https://doi.org/10.1007/s11042-012-1132-y.

A survey on visual adult image recognition

Christian X. Ries · Rainer Lienhart

Abstract We provide an overview of state-of-the-art approaches to visual adult image recognition which is a special case of one-class image classification. We present a representative selection of methods which we coarsely divide into three main groups. First we discuss color-based approaches which rely on the intuitive assumption that adult images usually feature skin-colored regions. Different ways of defining skin colors are described and example classification frameworks built on skin color models are presented. Another main group of approaches to adult image recognition is based on shape information which usually also exploit color information to find skin-colored regions of interest. Color and texture features are often used to augment such shape features. Finally we introduce approaches based on local feature descriptors.

Keywords Visual adult image recognition · Content-based image filters

1 Introduction

With the rapid growth of the amount of images publicly available on the Internet the need for reliable image content recognition has risen. Besides the obvious necessity of methods for image search and ranking it is also important to recognize unwanted or offensive images in order to be able to filter out these images. The largest group of images on the Internet which people may find offensive are by far adult images

This work was funded by Advanced Swiss Technology Group (ATG).

C. X. Ries (✉) · R. Lienhart
Augsburg University, Universitätsstr. 6a, 86159 Augsburg, Germany
e-mail: ries@informatik.uni-augsburg.de

R. Lienhart
e-mail: lienhart@informatik.uni-augsburg.de

since more than 10% of all websites feature adult contents.[1] Especially minors should be prevented from watching adult images but grown-ups may also not want to be exposed to such images, for instance by results they receive from online image search engines.

Thus, a lot of efforts have been put into adult image recognition by the research community. In this paper we provide an overview of different ideas researchers have come up with and compare them against each other as far as possible.

However, in this survey we concentrate on visual aspects only. In other words we do not consider approaches which for instance are based on context information such as image captions or website URLs. If an approach serializes [14, 18] or fuses [16] both text-based recognition and visual recognition, only the visual aspects of this approach are considered here.

We try to cover as many different approaches as possible, thus not all papers were selected based on their impact on the field or the number of references. Also, we wanted the number of papers presented in each section to roughly reflect the frequency of publications featuring the respective idea.

Note that visual adult image recognition is a special case of one-class image classification, thus the methods presented in this paper are usually inspired by approaches to solving more general problems. In fact, many of the basic methods presented in this survey were not invented for adult image classification in the first place and can thus also be applied for different classes.

2 Overview

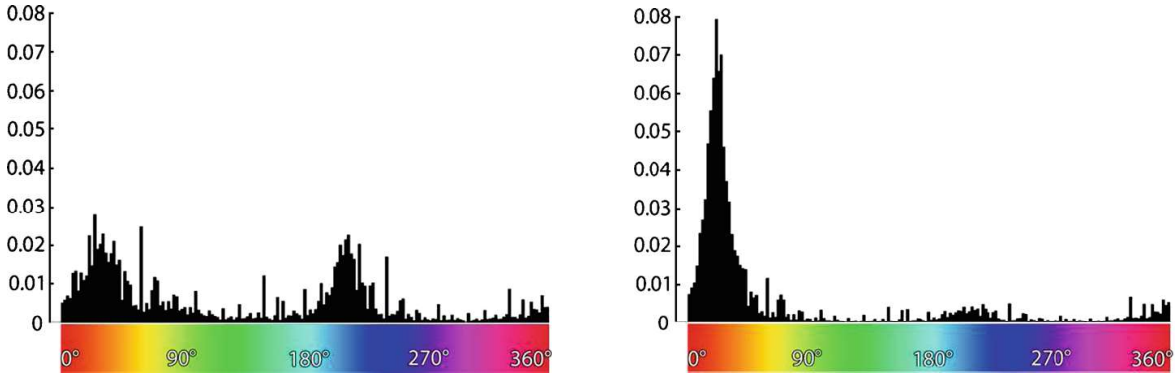
We discuss different approaches on visual adult image recognition. Due to the huge amount of work done in this field we focus on the main ideas and research directions as well as show frequently recurring approaches. For each of these main directions we present example approaches whereas we include the fundamental publications as well as some additional works.

We divide the current approaches into three main groups based on their underlying ideas. Note, however, that many researchers combine multiple ideas in their respective works.

The upcoming two sections explain adult image recognition based on color information. It exploits the assumption that adult color images usually feature large areas of skin colors. Due to the nature of adult content this assumption is likely to be valid for the majority of adult images (see Fig. 1). The second group of approaches which we present in Section 5 rely on shape information extracted from adult images. Some of both the color-based and shape-based approaches also include texture information. Texture however is commonly not used as a feature on its own and usually plays a minor role. Thus there is no section on texture-based recognition. Section 6 presents approaches which use local feature descriptors to recognize adult images.

Please note that throughout this survey approaches of different research groups are presented who usually use their own datasets for training and evaluation which

[1] According to http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html in October, 2011



(a) Flickr (b) Adult

Fig. 1 The hue value distribution of the pixels of **a** 100,000 random images downloaded from Flickr and **b** 100,000 images with adult content (including background pixels). The hue value is represented by an angle between 0° and 360° according to its definition in the HSV color space. The ratio of pixels in the skin color range (the *red-orange* area roughly with an extent of 30°) is notably higher for adult images

are not publicly available due to the nature of the subject. As a consequence, the individual results in most cases cannot be compared to each other quantitatively and we thus do not claim that approaches with better results are superior to others.

Further note that in most cases we use Receiver Operating Characteristics (ROC curves) in order to visualize the performances of various approaches. ROC curves plot the false positive rate (usually on the x-axis) against the true positive rate for different values of a parameter, usually a threshold of the classifier used. Despite careful extraction of the data points, the curves shown in this survey may slightly deviate from the original curves given in the respective publications.

3 Skin color definition

In the realm of color images the most obvious property of the majority of adult images is a large fraction of pixels which feature skin colors. For illustration, Fig. 1 shows the hue value distributions of the pixels of a set of 100,000 random images downloaded from Flickr and of a set of 100,000 images with adult contents. One can clearly see that the adult images are characterized by a higher ratio of skin-colored pixels.

On the one hand this fact can be exploited for building pre-filters or detectors for regions of interest. On the other hand color information can also be used for defining local or global image features based on which image classification is performed. This chapter introduces approaches for providing such image features based on skin colors and using them to recognize adult images.

All color-based approaches we found during our research for this survey require some kind of skin color model in order to determine whether a pixel is skin-colored. Therefore the next section describes the color models on which the approaches discussed in this paper are based.

There are many ways to define a skin color model. In this chapter which is based on the summary by Vezenevets et al. [34], we give a concise overview of some popular methods. It is important to note that many of the approaches mentioned here seem to

only consider skin colors of Caucasians. Often it is not mentioned how the problem of different skin colors is tackled or whether it is dealt with at all. One exception is Yang et al. [38] who state that skin colors of people of different ethnic groups mainly vary in terms of color intensity while the actual color tones are roughly the same. Others such as Yang and Ahuja [39] or Zheng et al. [44], however, feel that such differences should be modeled explicitly by mixture models.

3.1 Color intervals and constraints

The most straightforward approach to defining whether a pixel is skin-colored or not is to manually define a skin color interval in a color space which represents the actual color tone of each image pixel by a scalar. For example, Liao and Liu [23] define an interval for pixels in the HSV color space by simply setting up two boundaries on the hue scale in order to detect swimmers in a swimming pool.

In a similar way, Arentz and Olstad [1] define intervals on the Cb and Cr value of the YCbCr color space to find skin pixels. The interval borders are chosen empirically by analyzing the color frequencies of 500 images featuring skin pixels.

Duan et al. [10] also intersect two intervals which are defined on the I-component (i.e. red-orange tones) of the YIQ color space and the angle of the vector on the UV plane of the YUV color space, respectively.

An augmented version of the interval approach is to define the boundaries of a "'skin cluster"' in a multi-dimensional color space by using constraints. Kovač et al. [22] define such constraints for the RGB color space based on some properties of skin colors, e.g. minimum values for each color channel and constraints on inter-channel relations such as maximum difference among all channels. These constraints can be both constructed easily and evaluated rapidly.

The major drawback of these approaches is that the manual choice of a color space and the associated definition of the constraints is somewhat arbitrary and biased. Even on a one-dimensional scale it is not trivial to estimate the boundaries of the relevant interval by hand since skin may take on unexpected colors due to lighting conditions and reflections. Figure 2 illustrates this fact by showing the HSV color scale and a manually selected interval of non-skin colors. Only pure blue and green colors can be safely omitted when faced with noisy images such as compressed video frames and inconvenient lighting conditions.

Gomez and Morales [13] try to solve this problem by using a constructive induction algorithm which determines a decision rule automatically using an evaluation set of more than 32 million pixels. The resulting decision rules consist of conjunctions of constraints in the RGB color space. The best rule recognizes skin pixels with a precision of 93.6% at a recall of 94%.

Fig. 2 The hue values (of HSV color space) where a manually selected interval of non-skin colors in swimming videos is highlighted by a *red border* around the respective colors. The size of the remaining interval reflects the difficulty of defining a skin color interval manually. Image is a reworked version of an image from [29]

3.2 Color histograms

Another way to define a skin color model is to compute a skin color histogram and a non-skin color histogram from training images where pixels are manually labeled as 'skin' or 'non-skin'. This allows to compute a probability value (usually based on relative occurrence frequencies) for each color bin that colors falling into that bin originate from pixels depicting skin. With Bayes Theorem, the probability of a color being a skin color can thus be computed.

For example, Jones and Rehg [18] use a set of 13,640 images where they manually labeled skin pixels (in a subset of 4,675 images containing skin) to construct a 3D histogram with 256 bins for each color channel. In the same work, however, they show results suggesting that histograms with 32 bins for each color are sufficient and even slightly outperform the more detailed histograms at classification experiments since the latter tend to cause over-fitting.

Zheng et al. [44] use a similar approach. However, they create multiple histograms for different image clusters of brightness and chromaticity. On their test database of 138 adult images their approach produces a slightly better ROC curve for skin pixel detection than the before mentioned approach by Jones and Rehg [18]. Figure 3 shows a comparison of both approaches on the dataset provided by Zheng et al.

In a similar manner Zeng et al. [41] compute RGB histograms for skin color detection for three different brightness clusters corresponding to bright, normal and dark images, respectively, based on the average image brightness. Their results are also compared to the color model by Jones and Rehg [18] in Fig. 3 on skin pixels which have been manually annotated in 829 adult images.

Yang et al. [37] combine histograms based on three different color spaces. First they transform the RGB values of millions of skin pixels to two-dimensional values in order to construct a 2D skin color histogram and then delimit the skin color area by a closed curve. For each bin within the curve, a one-dimensional color value is computed based on color channel relations. To these values a standard relative frequency-based histogram approach is applied.

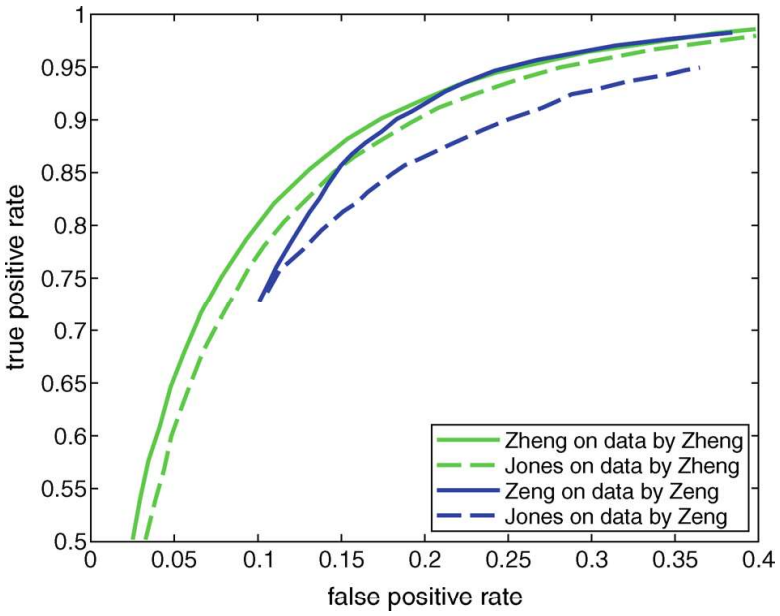


Fig. 3 The results of Zheng et al. [44] (*green curve*) at classifying skin pixels. They compare their approach to the one by Jones and Rehg [18] (*dashed green curve*) on their own dataset. The same comparison is done by Zeng et al. [41] (*blue curves*) on their dataset. *ROC curves* reproduced from [44] and [41], respectively. Comparison is only meaningful between curves of the same color

Like color constraints, color histograms can also be evaluated fast. Also, they usually can be easily updated with new training data. However, they sometimes lack the ability to generalize beyond training data which is limited to a finite number of depicted people and lighting conditions. The parametric approaches which are discussed in the upcoming section try to overcome this problem.

3.3 Parametric color distribution functions

Instead of using a discrete histogram to define the probability of each color to be caused by skin pixels, one can also derive a parametric function to model skin pixel distributions across a color space. Obviously, compared to histograms, parametric functions not only need much less storage space but also possess the ability to generalize by interpolating data points in the color space. The estimation of appropriate parameter values, however, can be complex and time-consuming. Therefore it usually cannot be performed on a very large set of training examples.

Hu et al. [16] try to alleviate this problem by estimating model parameters from a multidimensional histogram which they compute from a large set of pixels. Thus they try to combine the histogram approach with a parametric model.

A very popular parametric function to model skin color distributions is the Gaussian function. Yang et al. [38] for instance propose a 2D Gaussian on the plane defined by relative green and red values of RGB space. The mean and covariance of the Gaussian are computed from a relatively small set of human faces.

Yang and Ahuja [39] state that even though skin color pixels of different ethnic groups form a small cluster in the RGB and HSV color spaces, one multivariate Gaussian may still not be able to capture the real skin color distribution. They therefore use a Gaussian Mixture model consisting of two mixture components to define skin colors in the CIE LUV color space. The lightness value is discarded to reduce the lighting dependence of the color values. The parameters of the mixture components are estimated using the expectation maximization (EM) algorithm based on 2,447 of human faces.

Such Gaussian Mixtures are a popular approach to defining skin colors since many researchers second the opinion that the distribution of skin colors has multiple peaks in color spaces. Mixture models are for instance also used by Hu et al. [16], McKenna et al. [28] and Jebara and Pentland [17]. Jones and Rehg [18], however, state that in their experiments their histogram model outperforms the mixture models of the latter approach presumably due to the lack of degrees of freedom of parametric models.

In [16], Hu et al. examine how the number of Gaussian Mixture components influences the performance of the color model. Interestingly, they find that the performance of mixture models improves as the number of components is increased from one to five. For more components the performance levels out and even decreases when more than ten components are used due to over-fitting. This behavior can be observed for different color spaces. However the number of mixture components most likely depends on the diversity of training images.

A different parametric approach is used by Zheng et al. [42] who estimate the probability of the pixels of a given image depicting skin by using a maximum entropy model. They determine the probability distribution with the highest entropy with respect to constraints derived from training data. Since each possible color represents

one constraint for their maximum entropy model, the number of parameters is huge. Therefore, the parameters are estimated using Bethe trees which are graph structures capable of simulating pixel neighborhoods in a parameter free manner. Zheng et al. compare their model to the histogram model of Jones and Rehg [18] and achieve a slightly better ROC curve on the same database.

4 Skin color based adult image recognition

Once a skin color model has been defined, one can classify each pixel of an image as skin-colored or non-skin-colored. Based on this information several adult image recognition systems have been built.

All approaches described in this section essentially follow the same basic algorithm. A color model is applied in order to find skin-colored pixels. Based on this information, global image features are computed and used for traning a classifier, such as support vector machines (SVM) [10, 31], multi-layer perceptrons (MLP) [42], or decision trees [18]. In most cases, the features are statistical values such as the ratio of skin pixels.

4.1 Color-based global features

As mentioned above, Duan et al. [10] use the intersection of two color intervals to determine whether a pixel is skin-colored. They then apply an empirically determined threshold to the ratio of skin-colored pixels in order to separate adult images from non-adult images. On a test set of 312 adult images and 710 non-adult images, this filter yields 93.91% true positives at 29.86% false positives.

All images which pass the color filter are then classified using a Support Vector Machine which is trained on the color histograms and color coherence vectors of the training images. Combining the results of the pre-filter and the SVM classification yields a total true positive rate of 80.7% at a false positive rate of 10%.

Jones and Rehg [18] compute five different features for each image based on skin-colored pixels identified by their skin color histogram model such as the percentage of skin pixels or the number of connected skin pixel areas. They train a C4.5 decision tree on these features which yields surprisingly good results on 5,453 adult images and 5,226 non-adult images. They obtain the ROC curve shown in Fig. 4 which features for example 85.8% true positives at 7.5% false positives. Similar basic features are used by Zheng et al. [42, 43]. They, however, also fit global and local ellipses to the skin pixels to obtain further features such as the relative number of skin pixels within these ellipses or features relating to the size and orientation of the local ellipses. A MLP is then trained on the features of 1,297 adult images and 3,787 non-adult images. The ROC curve obtained from a test set of the same size is shown in Fig. 4 along with the ROC curve of Jones et al. on their respective dataset.

Rowley et al. [31] use 19 different features for adult image recognition, nine of which include color information. Based on the color model of Jones and Rehg [18] they generate a skin color map for the image and determine connected components in the map. The number and compactness of these components constitute features as well as the means and standard deviations of the full skin map and the skin map of the connected components. Rowley et al. also use the ratios of edge pixels within skin

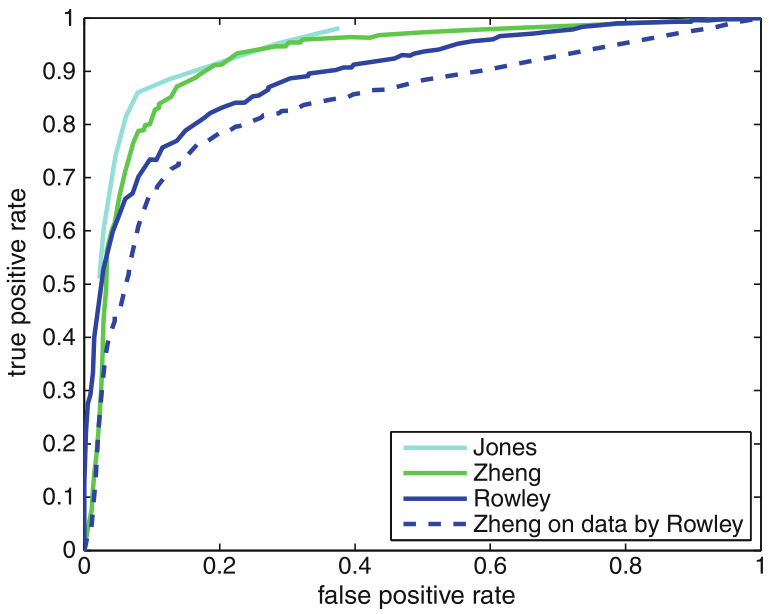


Fig. 4 The ROC curves of the adult image recognition systems of Jones and Rehg [18] (*red curve*), Rowley et al. [31] (*blue*) and Zheng et al. [42] on their own dataset (*green*) and on the dataset of Rowley et al. [31] (*dashed blue*). Note that different colors represent different datasets. The data in this figure was extracted from the figures featured in [18, 42] and [31]. Also note that the curves were combined into one figure only for the sake of brevity. Comparison between different datasets is meaningless

regions as features, since there is usually little texture on human skin. Straight lines are detected as a final feature since they are unlikely to appear on human bodies. An SVM is trained on these features computed on 812 adult images and 16,488 non-adult images. The SVM yields the ROC curve also shown in Fig. 4 on a test set of 1,331 adult images and 50,629 non-adult images.

4.2 Combined color features and face detection

One common problem that often occurs with color-based adult image detection is a considerable confusion rate of adult images with close-up face images. Choi et al. [6] tackle this problem by a two step framework. First they train a multi-class support vector machine on MPEG-7 visual descriptors of close-up faces, adult images and non-offensive images. Other approaches using MPEG-7 descriptors for adult image recognition will be discussed in Section 5.6. Since the actual final classifier is based on color, we list this approach among color-based recognition.

The SVM is trained on three classes (adult, non-adult and faces) in order to obtain an output weight for the classes “adult” and “non-adult”. In the second step of the framework the non-face images are classified by a Bayesian decision rule which multiplies the output weight of the multi-class SVM with a posterior probability of a given image being adult. This probability is simply derived from a Gaussian distribution of skin pixel ratios over training sets of adult images and non-offensive images. For computing these ratios, Choi et al. use the color histogram model of Jones and Rehg [18] to determine whether a pixel is a skin pixel. Results in terms of true positive or false positive rates are not given.

4.3 Summary on color-based adult image detection

The color-based approaches presented in this section all use a color model which detects skin-pixels and then use this information for devising image features. The approach by Choi et al. [6] also integrates face detection in order to reduce the apparently high confusion rate with close-up face images.

Table 1 Summary of the color-based approaches described in Section 4

	Color model	Features	Classifier	# pos	# neg	TP	FP	Remarks
Duan et al. [10]	Intervals	Skin ratio, histogram, coherence vector	Threshold/thresh. + SVM	312	710	93.91/80.70	29.86/10.00	
Jones and Rehg [18]	Histogram	Skin ratio, color probabilities, connected components	C4.5 decision tree	5453	5226	85.80	7.50	
Zheng et al. [42]	Max. entropy	Skin ratio, skin areas, skin ellipses	MLP	1297/1331	3787/50629	80.00/77.00	8.00/20.00	Approx. TP and FP taken from ROC curve, second numbers on data of [31].
Rowley et al. [31]	Histogram	19 features (statistics, connected components)	SVM	1331	506929	84.00	20.00	Approx. TP and FP taken from ROC curve.

For each approach, the color model, the features used and the classifier are listed. The true positives (TP) and false positives (FP) are given for each approach as well as the respective number of adult (# pos) and non-adult (# neg) test images

The features used by the different approaches are relatively similar as they all compute pixel-based statistics to some extent. However, the number of features used varies greatly from one feature [10] to 19 features [31]. Also, different classifiers are used such as SVM and MLP. Duan et al. [10] examine the impact of increasing the number of features and using an SVM instead of simple thresholding. Even though they observe a better classification performance, the difference is not as significant as one would expect.

Table 1 provides an overview of all approaches presented above. Unfortunately, the significant differences between the respective test sets prohibit direct comparison (the sizes of test sets vary by a factor of up to 50).

The main advantage of color-based classification is that features can usually be computed quickly and implementation is often not very complex. Color-based classifiers apparently also yield a relatively high true positive rate. However, they usually suffer from high false alarm rates, since skin-like colors appear on many natural objects such as sand and wood. Also, some non-adult images of people feature skin color distributions which are probably almost identical to those of adult images, for instance people in swimwear.

Another obvious disadvantage is that color-based approaches cannot be used on grayscale images or on images where there is not much visible skin despite featuring adult contents, e.g. because the depicted people might still be clothed to some extent or the adult content might be part of the background. The latter kind of images, however, is hard to catch by any approach since the adult contents are limited to subtle details in such images.

As mentioned in the introduction to this chapter, another useful application of information obtained from skin color models is filtering out image background or determining regions of interest for further processing. The methods described in the next chapter only operate on image regions featuring skin colors after applying one of the color models explained above to determine such regions.

5 Approaches including shape information

Besides often featuring a distinct color distribution most adult images also share some characteristic shapes. Therefore, many researchers formulate features based on shapes present in adult images. For example, as mentioned in the previous section, Zheng et al. [42] create among other (color-related) features a global and one or more local fit ellipses around the skin-colored pixels inside of each image. One could argue that this is a way to include basic shape information, since the ellipses describe the rough boundaries of the skin-colored areas. In a similar way the compactness of a skin-colored region which is used as a feature by Rowley et al. [31] can also be considered basic shape information.

All shaped-based approaches which are presented in this survey also rely on skin-colored pixels to a certain degree. Thus, their performance highly depends on the quality of the underlying color model and they obviously have to deal with the same problems regarding skin detection as described above. Unfortunately, the impact of skin detection accuracy is not explicitly evaluated by the authors of the works presented here. Also note that some approaches include texture-based features which are, however, usually closely related to either shape features or color features.

Again, all approaches described in this chapter share a similar basic framework. First, skin-colored regions are determined using a color model. Afterwards, the shape of these regions is described by a number of features which form a global description vector for the respective image. We divide the shape features into five groups: contour-based features [1, 16], color segments [4], geometric constraints [11], moments [19, 35, 41, 44] and MPEG7 features [21, 40]. Finally, the features are used to train a classifier (in most cases SVM) or for *k*-nearest neighbors classification.

5.1 Contour-based features

Shape information is for instance used by Hu et al. [16], Yang et al [37] and Wu et al. [36] (all three publications share at least one author). They first divide images into regular rectangular blocks and then determine whether the fraction of skin pixels within the individual blocks surpass a threshold and thus obtain connected regions of skin blocks.

The inner corners of the blocks of the largest connected region are then considered the points of interest from which they start searching for the contour of the actual skin-colored object. They connect the points of interest to obtain a closed curve which is reduced to the inside by shifting its points orthogonally inwards until a skin pixel is found. Eventually they obtain an outline of the largest skin-colored object within the image.

They then compute features based on the relative positions of large non-skin regions within this outline (modeling underwear or swimwear). Also, the middle curve of the skin region is determined which allows a more expressive (i.e. relative) description of these positions. The aspect ratio of the image is included as a feature as well. The resulting feature vectors are classified by determining the class of the respective nearest neighbor among the training examples in feature space. According to Hu et al. [16] they obtain a true positives rate of 92.8% on 1,000 adult images at a false positive rate of 6.0% on 1,000 non-adult images using the same amount of images from each class for their nearest neighbor search.

Unfortunately, the authors do not comment on the runtime of their approach. The process of determining the exact outline of the skin-colored blob includes some complex operations such as pixel-wise narrowing down the contour.

In [36] a color-based pre-filter is added to the framework which is designed to detect human presence in an image. It is based on the connected skin block regions which are devised as explained above and yield an image description which is a set of skin areas characterized by their numbers of blocks. A probability model is created for these sets by error minimization. Thus, each image is assigned a probability of featuring people which is compared against an empirically determined threshold. On a different dataset, Wu et al. show that this pre-filter approximately halves the number of false positives and considerably speeds up the visual recognition process.

Arentz and Olstad [1] also use a set of features including contour information about skin-colored objects. They trace the outline of the skin-colored objects (found by applying color intervals in YCbCr color space) and compute the distance to the object's centroid at each point of the contour. The sequence of distances is then normalized and the Fourier Transform is applied to obtain a fixed number of energy-coefficients which constitute the feature vector. The relative size of the object and its position in the image is also added to the vector.

Besides shape features Arentz et al. also include histograms of the Cb and Cr values found in the image in their feature vector. Finally they add a texture-based feature, which is basically a histogram of color differences between neighboring pixels.

The feature vectors of 365 adult images and 575 non-offensive image are then used to train a genetic algorithm. A test set of 500 adult images and 800 non-adult images is used to evaluate the algorithm's performance. The true positive rate of the genetic algorithm is 92.1%. On the non-adult images a false alarm rate of 10.6% is achieved whereas on a subset of the non-offensive images showing only portraits, 26.5% are classified as adult images. The reason for this difference is the fact that portraits usually feature large areas of skin and occasionally shapes which can also be found on human bodies.

5.2 Moments

Moments are a way to express the shape of a set of points (e.g. a set of skin-colored pixels) mathematically and are thus a popular means of quantizing shapes. In this context, probably the best known features are the normal moment invariants introduced by Hu [15]. Hu points out that each density function (e.g. the skin pixel distribution of an image) can be represented by its two-dimensional moments. Thus, skin color distributions can be characterized by a set of moment invariants independently of the skin region's position and size in the image. These invariants correspond for example to the 'spread' or 'slenderness' of the skin pixels. In a very similar way, Zernike moments [7] model the shape of a distribution.

5.2.1 Hu and Zernike moments

Zheng et al. [44] suggest an approach which uses both Hu and Zernike moments. They first use their clustered color histograms (see Section 3.2 for details) to identify skin regions within images. The skin regions are refined by morphological operations in order to obtain a homogeneous segmentation.

Along with the Hu and Zernike moments of the skin distribution, straightforward general shape descriptors are included in the feature vector based on skin pixel ratios. These shape features are then computed on 897 adult images and 732 non-adult images. The resulting sets of feature vectors are used to train AdaBoost classifiers based on various weak classifiers, namely stumps, C4.5 trees, SVMs and MLPs.

According to the experiments presented, C4.5 trees yield the best classification performance of 89.2% true positives at 15.3% false positives which are the average values from a five-fold validation. These results are roughly on par with most color-based approaches described in the previous chapter. One reason for this fact might be that all shape features computed by Zheng et al. are solely based on the skin-pixel distribution.

Apparently the same features[2] which are used by Zheng et al. are used by Ka [19] to describe differently rotated versions of manually annotated obscene objects (i.e. body parts usually exposed in adult images only). SVMs are trained on these feature

[2]The set of features used in [19] is not further specified. However, Zheng et al. [44] and Fleck et al. [11] are cited, therefore moments are presumably used.

vectors and used to classify the contents of multi-scale sliding windows on skin-colored areas of query images.

Ka evaluates this method on various test sets of different difficulty, ranging from landscape images to "slightly adult" images. The test sets also vary in size between 20 images and 2,000 images. According to the numbers given, Ka's approach yields a true positive detection rate of about 91% to 93.6% at a false alarm rate between 0% and 7% depending on the difficulty of the test set. Note that subjectively ambiguous images apparently are always counted as correct classifications.

5.2.2 Moments on edge maps

Moment-based features are also used by Wang et al. [35]. They, however, compute the moments on an edge image obtained by Daubechies' Wavelet transform. Since Wang et al. aim for an efficient approach, they compute these relatively expensive features only for images which pass a fast to compute color filter: Only if an image passes an efficiently implemented color histogram comparison, the moment features are computed. For classification a nearest neighbor search is performed in feature space among 500 moment features of adult images and 8,000 features of non-adult images. Wang et al. tested their system on a set of 437 adult images and 10,809 non-adult images and obtained 97.5% true positives at 18.4% false positives or 92.5% true positives at 10.7% false positives respectively depending on the threshold settings.

5.2.3 Combined moments, color and shape

Like Zheng et al. [44], Zeng et al. [41] also use Zernike moments in combination with the area of the skin-colored region of an image and the edge intensity of the respective region as shape-features for adult image recognition. They also include color features such as mean and variance of skin-colored pixels in the respective image. Additionally, they compute texture features, namely the texture contrast and the texture coarseness.

The classification performance of C4.5 trees and two different SVM kernels are compared on a test set of 11,349 adult images and 59,057 non-adult images. The best results are achieved when using a RBF kernel SVM. It yields 76.5% true positives at 5% false positives. As mentioned in Section 3.2, the framework of Zeng et al. employs the color model of Jones and Rehg [18] which is augmented by devising three different histograms for dark, normal and bright images, respectively. This color model is used to find skin regions from which all features are computed. The skin regions are refined by eliminating small skin regions and skin regions featuring a relatively high color variance since such regions are considered to be textured too much to be human skin.

Note that this approach by Zeng et al. combines features based on color, texture and shape. Features of all three of these types are also defined by the visual portion of the MPEG-7 standard which yields the features used by the approaches introduced in the next section.

5.3 MPEG-7 descriptors

MPEG-7 is a popular standard for describing multimedia content. It defines low-level descriptors for both audio and visual data. The visual descriptors include among

others color, shape, and texture descriptors, and are often used as features for image content analysis.

A detailed overview of the MPEG-7 color descriptors is provided in [27]. All five color descriptors are based on a color histogram. One color descriptor, for instance, models the dominant color of the image and the color layout. Another color descriptor, namely the Color Structure Descriptor, also includes basic shape information since it encodes the spatial color distribution.

However, MPEG-7 also defines distinct shape descriptors: a region-based descriptor and a contour-based descriptor. For more details on the MPEG-7 shape descriptors the reader is referred to [3].

The MPEG-7 texture descriptors include an edge histogram (which also models the spatial distribution of shapes to some extent), a homogenous texture descriptor and a texture browsing descriptor. The texture descriptors are explained in detail in [27].

The following approaches use all of these groups of visual MPEG-7 descriptors and are therefore only partially shape-based. However the use of shape-related descriptors is the main difference from the approaches mentioned in Section 4, since the MPEG-7 color descriptors are similar to some of the color features explained in Section 4.

Yoo [40] for instance uses three of the MPEG-7 visual descriptors, namely the edge histogram, the color layout descriptor and the homogenous texture descriptor. These features are computed for three groups of adult images and a set of non-adult images and stored in a database using an efficient indexing structure. This database is then used for k-nearest neighbor search. Among several experiments, the highest true positive rate of 99.25% is obtained when the database contains a total of 7,500 adult images and 1,300 non-adult images. The respective false alarm rate is 23.00%. Increasing the number of non-adult images in the database to 10,000 reduces the true positive rate to 90.96% and the false alarm rate to 5.75% (on a different, larger test set of 2,000 instead of 400 images, however).

In another approach based on MPEG-7 descriptors, Kim et al. [20, 21] use the same descriptors as Yoo plus color structure, region shape and dominant color (omitted in [20]). In [21], they train one artificial neural network for each MPEG-7 descriptor separately (presumably on about 1,300 adult images and 1,300 non-adult images). They evaluate their neural networks on different test sets, i.e. the number of evaluation images varies between 3,803 and 5,635. The ratio of adult images to non-adult images also varies significantly.

The network which is based on color structure performs best yielding a true positive rate of 94.69% at a false alarm rate of 4.9%[3] on 2,694 adult images and 2,703 non-adult images. Note that not only are the sizes of the test sets different for each descriptor, but the color structure also has by far the highest number of dimensions. Therefore it appears to be more expressive and better suited for the authors' neural network configuration of 2×10 hidden nodes than the remaining descriptors.

In [20], Kim et al. use a similar approach to distinguish normal images from images of people wearing swim suits and three different levels of adult images. Since they try

[3] In the text of Kim et al. [21] the number of false positives is stated to be below 3%. Judging from the given numbers, however, this appears to be the value of false positives computed against the total amount of images.

to distinguish five different images classes, they use a multi-class neural network. The training set consist of 1,702 images for each class and the test set of 1,700 images (the sizes of the individual classes are not given). Surprisingly, this time the homogenous texture-based network performs better than the networks based on color structure and edge histograms.

The overall best result is achieved when using the combination of color layout and homogenous texture where the true positive classification results for the three classes of adult images range from 78.53% to 87.65%.

Note that the three adult classes are often confused with each other. Since we are interested in a binary result for this survey, we ignore confusion among these three classes and thus deduce an overall average true positive rate of 91.37%. The false alarm rate is 0.59% for swim suit images and 7.35% for non-adult images yielding a total false alarm rate of 7.94%.

5.4 Color segments

Bosson et al. [4] create features which describe the shape of skin-colored image segments. They use a HSV histogram to find initial skin segments and grow them to include less likely skin pixels in the vicinity using empirically determined thresholds (similar to Canny's edge detector [5]). Their approach is tested to classify 82% of all skin pixels in 1,000 test images correctly at a false alarm rate of 17%.

For each image a number of shape features is computed based on the identified skin regions: the fractional area of the largest skin region, the number of skin segments (a segment being a connected area of pixels whose color falls into the same skin color bin), the fractional area of the largest skin segment and the fractional area of skin that is enclosed by the result of a commercial face detector. Since real-world photos are considered to feature a higher diversity of colors than for instance graphics, the number of different colors is also added to the feature vector.

Four different classifiers are compared by Bosson et al. on various test sets. The combined test sets contain 3,967 adult images and 7,038 non-adult images of varying difficulty (from showing people to graphics). The classifiers used are a generalized linear model, the *k*-nearest neighbor algorithm, MLP and SVM. Figure 5 shows the respective ROC curves. The *k*-nearest neighbor algorithm yields the least number of false positives, the MLP and SVM on the other hand perform better regarding the true positive rate whereby the MLP slightly outperforms the SVM. According to the paper, for standard parameters the MLP yields 83.87% true positives at 10.86% false positives while the *k*-nearest neighbor algorithm yields 84.57% true positives at 11.57% false positives.

5.5 Geometric constraints

A different approach to using shape for adult image recognition is proposed by Fleck et al. [11]. Their approach is based on a geometric human body model which is defined by a set of constraints.

They also start by finding skin-colored regions. For this purpose, they transform each pixel's color value into log-opponent representation as described in [12] yielding an intensity value and two hue values. They then apply decision rules in order to

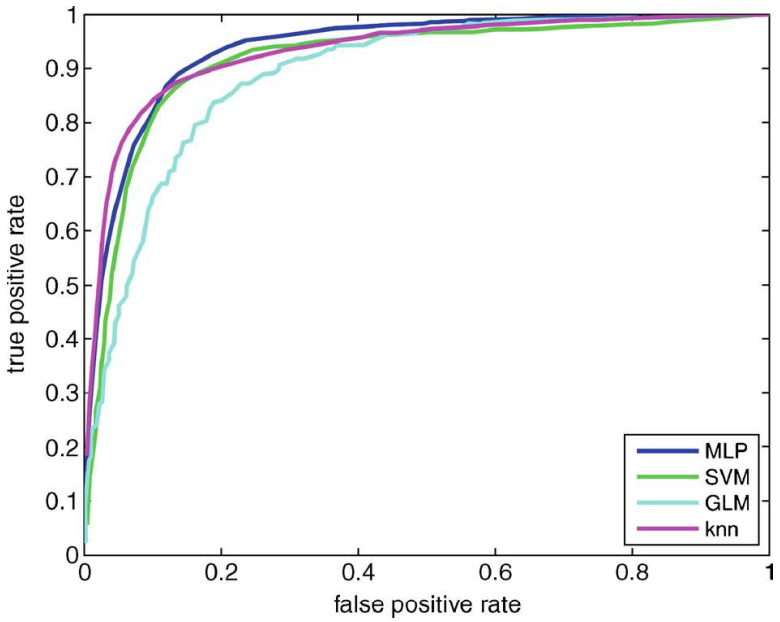


Fig. 5 The ROC curves for the four classifiers used by Bosson et al. in [4]: Generalized linear model (*GLM*), multi-layer perceptron (*MLP*), *k*-nearest neighbors (*knn*) and support vector machine (*SVM*)

find skin-colored pixels. Additionally, like Zeng et al. [41], Fleck et al. also consider regions with high texture as non-skin regions.

After identifying the skin-colored regions of an image, Canny's [5] edge detector and Hough transform are used to find roughly straight near-parallel pairs of lines which are candidates for human limbs. These candidates are iteratively combined according to a set of constraints which model human body geometry. If it is possible to assemble the limbs in a geometrically reasonable way, the image is classified as adult (since all assumed body parts are skin-colored). Fleck et al. test their method on 138 adult images and 1,401 non-adult images (featuring a large variety regarding quality and content). They obtain a true positive rate of 52.2% and a false positive rate of 3.4%.

Interestingly, Fleck et al. also state results for standalone usage of their skin filter which classifies 87.3% of their adult test images and 7.4% of their non-adult test images as adult. Thus, their geometric analysis eliminates more than half of the false positives and about one third of the true positives.

5.6 Summary on shape-based approaches

As the large number of shape-based approaches selected for this survey reflects, shape features are a very popular feature type for image classification. However, many of the features mentioned above model similar characteristics which introduce a high level of abstraction. This abstraction is introduced due to the vast variety of possible complex poses of human bodies. For training standard machine learning algorithms, the poses are reduced to some common characteristics which can roughly be reproduced for most plausible human poses. However, this also means that usually much shape information is omitted and not included in the respective model. To our knowledge, no approaches exist which explicitly try to model articulated human poses for adult image recognition in terms of the recently popular body part-based models for example.

Table 2 Summary of the shape-based approaches described in Section 5

	Features	Classifier	# pos	# neg	TP	TP	Remarks
Hu et al. [16]	Features based on skin-colored region of interest	1-nearest neighbor	1,000	1,000	92.80	6.00	
Arentz and Olstad [1]	Coefficients of Fourier Tranform on distances of oultine pixels of skincolored area to center pixel	Genetic algorithm	500	800	92.10	10.60	
Zheng et al. [44]	Eccentricity, compactness, rectangularity, Hu moments and Zernike moments of skin region	AdaBoost with C4.5 trees	1,331	50,629	84.00	20.00	5-fold cross validation.
Ka [19]	Moments on body parts	SVM	2,000	240	93.60	7.00	Subjectively ambiguous images did not affect the classification rates.
Wang et al. [35]	Moments on edge image	*k*-nearest neighbors	437	10,809	92.05	10.07	
Zeng et al. [41]	Zernike moments, skin area size, edge intensity	SVM	11,349	59,057	76.50	5.00	
Yoo [40]	MPEG7 visual descriptors	*k*-nearest neighbor	2,400	2,000	9,096	23.00	
Kim et al. [21]	MPEG7 visual descriptors	Neural networks	2,694	2,703	94.69	4.90	
Bosson et al. [4]	Features derived from skin segments	SVM/MLP	3,967	7,038	83.87/84.57	10.86/11.57	
Fleck et al. [11]	Near-parallel pairs of lines	Threshold	1,401	138	52.20	3.40	

For each approach the features and the classifier are listed. For more details on the numerous different features please refer to the respective section. The true positives (TP) and false positives (FP) are given for each approach as well as the respective number of adult (# pos) and non-adult (# neg) test images

Besides methods for finding regions of interest and using their contours for feature computation, the approaches presented in this chapter also include relatively uncommon features such as geometric constraints and color segments. The majority of approaches, however, use more conventional shape features such as moments or MPEG-7 descriptors. Also, different popular classifiers have been used throughout the referenced works, such as SVM, MLP and *k*-nearest neighbors. None of them, however, seems to clearly outperform the others (as far as several classifiers have been tested by the authors). The choice of the actual classifier thus apparently plays a minor role.

It should also be emphasized that all of the approaches discussed above implicitly assume that we know for each pixel of an image whether it is skin-colored. Thus, there is not a clear distinction between color-based and shape-based approaches. Also, the performance of shape-based approaches highly depends on the quality of the skin-color detector used. The color model defines the regions of interest whose shapes are then translated into feature descriptors. For the same reason it is not surprising that the results of shape-based approaches seem to be slightly better on average than approaches which only use color.

A summary of the shape-based approaches is given in Table 2.

6 Approaches based on local features

The extraction of local features from images is a popular approach to object detection and recognition. A large number of different local feature descriptors have been proposed by the research community and even more modifications of existing ones have been suggested.

Local feature descriptors describe an image region by means of a feature vector. Hence they transform an image region surrounding a given pixel position (commonly called a key point) into a vector with fewer dimensions than the actual image region (with regard to the number of pixels), while still providing a general appearance description which is similar for similar regions. The key points are either chosen by an interest point detector which is called sparse sampling or on a regular grid which is known as dense sampling.

One of the best known feature descriptors is for instance SIFT (scale invariant feature transform) [26] which is usually combined with a Difference-of-Gaussian (DoG) key point detector. SIFT has also been used for adult image recognition as well as Self-Similarity [32] and PCA and DCT transformed image patches.

It is usually an important advantage of local feature descriptors that they can be computed independently from color information. However, the approaches described below still include some color information during feature computation. Given the task of adult image recognition it is reasonable not to omit color information.

Another advantage is that local feature descriptors expressively yet compactly model the appearance of image regions by a fixed-sized vector representation which facilitates the comparison of image regions and thus images. Also, different types of local features are usually easily interchangeable for experiments. However, computing local features is many cases more time-consuming than examining e.g. color-related image features.

6.1 Bag-of-features

The basic algorithm for bag-of-words image classification consists of three major steps. First, visual features are extracted from images. To each visual word a cluster id is assigned which is obtained from a previously performed clustering on a large set of features. The clusters are also known as visual words and the set of clusters is called a visual vocabulary. Most commonly, *k*-means clustering or a hierarchical variant of *k*-means is used. For each image, counting the numbers of occurrences for each visual word finally yields a word-occurrence histogram which is also called a bag-of-(visual-)words histogram.

These histograms are vector representations of the respective images and can be used to train classifiers, such as SVM.

6.1.1 Bag-of-features on SIFT

In [25] Lopes et al. apply a bag-of-features approach to adult image recognition. They extract local feature descriptors on sparse key points. Since they use SIFT as local feature descriptor, they supposedly use a DoG key point detector. They compare SIFT computed on grayscale images to SIFT computed on the Hue value of the HSV color space (called Hue-SIFT). This way Lopes et al. include color information in their feature descriptors as most adult images contain skin-colored pixels.

Lopes et al. train a linear SVM on their bag-of-words vectors using five-fold cross validation on a test set of a total of 180 images (the amount of adult images and non-adult-images is not specified). The resulting SVM yields a maximum correct image recognition of 84.6%. Interestingly according to the results given in the paper the Hue-SIFT model considerably outperforms a standard SIFT model. Figure 6 shows the recognition performance plotted against the size of the respective visual vocabulary. The latter parameter, however, appears not to affect the performance significantly.

6.1.2 Bag-of-features on transformed image patches

Deselaers et al. [9] also propose a bag-of-words approach for adult image recognition. However, they prefer the PCA transformed image patches (also around Difference-of-Gaussian key points) as features since the PCA reduces the dimensionality of the respective image patch while preserving color information. For classification a SVM is used.

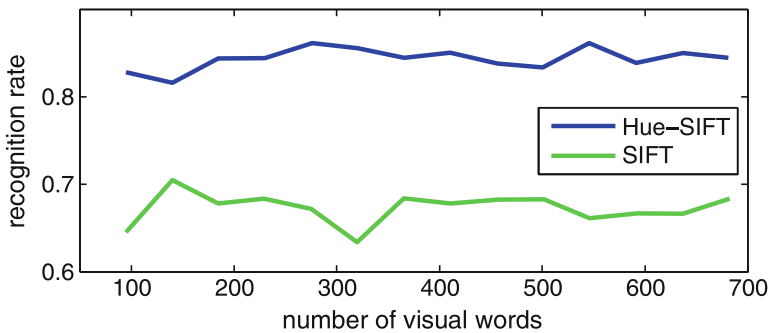


Fig. 6 The recognition performance of both the SIFT and the Hue-SIFT bag-of-features models used by Lopes et al. [25] plotted against the size of the respective visual vocabulary

They also choose a different clustering strategy which splits the feature vectors by fitting Gaussian mixture models into feature space. Deselaer et al. find that the performance of their system does not improve beyond 11 splits which correspond to 2,048 visual words.

For testing their approach, they use the same database as Yoo [40] on which they perform five-fold cross-validation. Deselaer et al. show results for different ground truth definitions of offensive images (which seem to deviate from the partitioning of Yoo). If they consider lightly dressed persons offensive (and thus positive examples), their classification result is 99.2% true positives at 0.5% false positives. If at least explicit nudity is required for defining an image as positive, they obtain 95.3% true positives at 3.4% false positives. On a significantly more difficult dataset which has been created from web images, the respective rates are 93.9% true positives at 17.2% false positives and 78.8% true positives at 9.4% false positives.

In a similar approach to Deselaers', Ulges and Stahl [33] also use bag-of-words methods in order to identify adult images. In contrast to the other publications listed in this survey, their work also includes experiments on images containing child sexual abuse (CSA) in collaboration with the police.

The features used by Ulges and Stahl are the low-frequency coefficients of the Discrete Cosine Transform of the respective image patch in YUV color space. These features are clustered into a visual vocabulary of 2,000 words.

For classification, Ulges and Stahl also use SVM. They compute results for five-fold cross-validation on 1,000 adult images and three different negative test sets with 1,000 images each. These negative sets are non-adult images from different sources such as Flickr or a web crawler. The results range from 90.3% true positives with 9.7% false positives to 94% true positives with 6% false positives, which corresponds to equal error rates of 9.7% or 6%, respectively.

Interestingly, the tests on CSA images yield significantly higher equal error rates between 11.2% and 21.8%. Another interesting experiment done by Ulges and Stahl is classifying the CSA images while using "normal" adult images as negative images. Surprisingly, they still obtain a fairly good result of 74% true positives and 26% false positives which is nearly comparable to one of their previous experiments, where the negative images are non-adult web images.

Ulges and Stahl also implement a color-based classifier analogous to Jones and Rehg's [18] for comparison. They find that their bag-of-words system outperforms the color-based classifier by up to 4, 4% for adult images and up to 18, 5% for CSA images depending on the negative dataset.

6.2 Probabilistic latent semantic analysis

Lienhart and Hauke [24] compute pLSA (probabilistic Latent Semantic Analysis) models to represent adult images and non-adult images. A pLSA model consists of a given small number of topics which describe the image content. Each topic causes a different distribution over the set of visual words. Therefore observing different combinations (co-occurrences) of visual words entails that certain topics are present in the image.

Thus, in this approach a low-dimensional topic vector is derived for each image which is computed from the respective co-occurrence vector. The co-occurrence vectors are simply histograms of visual words like in the above approaches. The

topics are modeled as latent variables and the topic distribution is derived using an Expectation Maximization algorithm given the visual word distribution over the images. As a result, a conditional probability distribution of topics given visual words per image is obtained. The pLSA transforms the co-occurrence vectors of query images into topic vectors according to this distribution. Intuitively, the combination of co-occurring words of an image is explained by a low-dimensional topic vector.

Note that interestingly this approach does not require the visual vocabulary to be computed from a set of images including adult images as described in [30].

Lienhart and Hauke use two different local feature descriptors: SIFT and Self-Similarity which they both compute on a dense grid. Both features yield comparable performances. The features are computed in different color spaces by concatenating feature vectors computed from each color channel. The classification performance remains roughly constant except for a few color spaces which seem not as suitable as the others. Figure 7 shows the results of the color space comparison for the slightly better Self-Similarity features. For classification Lienhart et al. perform a k-nearest neighbor search on a reference set of topic vectors computed from 600 adult and 7,076 non-adult images. Overall, the best configuration (Self-Similarity in HLS color space) achieves a false alarm rate of 1.9% on 12,511 non-adult images and a true positive rate of 92.7% on 2,068 adult images.

6.3 Summary on local feature-based approaches

Today local features are among the most popular methods of describing image content for classification and retrieval systems which is reflected by the huge amount on publications on this subject matter. For adult image detection, however, local features have still been used relatively rarely in comparison to color or shape-based approaches.

Still, approaches based on local features yield superior results compared to most color or shape-based approaches (on the datasets of the respective authors). Thus,

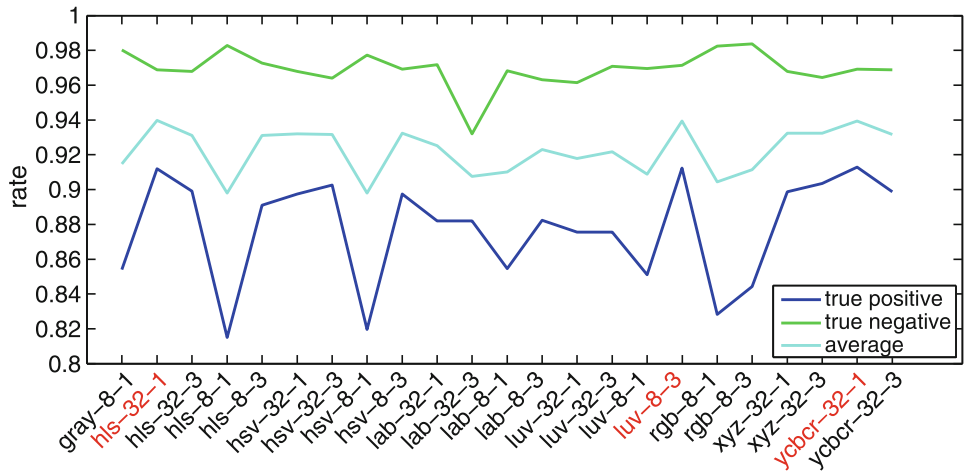


Fig. 7 The comparison of classification performances of pLSA models computed by Lienhart and Hauke [24] on Self-Similarity on different color spaces (encoded as ⟨color space⟩-⟨bits per channel⟩-⟨channels used for descriptor⟩). Results are averaged results from k-nearest neighbor searches with each $k \in \{5, 7, 9, 11, 13, 15, 17, 19\}$. The number of visual words is 500 and 50 topics are used. Figure reproduced from [24]. Best results are indicated by *red labels*

Table 3 Summary of the local feature-based approaches described in Section 6

	Features	Classifier	# pos	#neg	TP	FP	Remarks
Lopes et al. [25]	Bag of words: Hue-SIFT	SVM	<180	<180	84.60	na	5-fold cross validation. Ratio might be accuracy instead of TP.
Deselaers et al. [9]	Bag of words: PCA on image patches	SVM	~5100	~400	95.30	3.40	5-fold cross validation.
Deselaers et al. [9]	Bag of words: PCA on image patches	SVM	1000	1000	78,80	9.40	5-fold cross validation. Different dataset.
Ulges and Stahl [33]	Bag of words: DCT on image patches	SVM	1000	1000	90.30	9.70	5-fold cross validation.
Ulges and Stahl [33]	Bag of words: DCT on image patches	SVM	1000	1000	94.00	6.00	5-fold cross validation. Different negative set.
Lienhart and Hauke [24]	Bag of words: Self-Similarity → topic vectors by pLSA	*k*-nearest neighbors	2068	12511	92.70	1.90	

For each approach, the features and the classifier are listed. The true positives (TP) and false positives (FP) are given for each approach as well as the respective number of adult (# pos) and non-adult (# neg) test images

it appears to be a promising direction for future work in this area. It also seems reasonable to combine local features and bag-of-words models with global color or shape features, which are for instance used as pre-filters.

Another important insight is that including color information in the local feature descriptors apparently improves bag-of-words models for adult image detection. Furthermore, transforming the word histogram vectors into topic vectors by applying pLSA also helps improving the performance. This is probably due to the fact that the raw occurrence statistics are augmented by semantic concepts when applying pLSA.

All approaches (except for [24]) rely on SVM classification. Considering the previous chapters of this survey, however, it is reasonable not to expect major differences in performance for different classifiers.

Table 3 summarizes the approaches described in this chapter.

7 Summary and conclusion

In this survey we have given an overview of the state-of-the-art in visual adult image recognition which we have divided into three main groups: based on color, shape and local features. This section summarizes and discusses the approaches and gives an overview of the individual results.

7.1 Summary

Many researchers build features based on the color distribution of images, since one can expect a certain amount and a certain spatial arrangement of skin-colored pixels in most adult images. Finding these pixels, however, requires devising a color model beforehand which is difficult due to large variations in the appearance of human skin under different lighting conditions and of different races across the world. Thus many different approaches to defining skin color exist. The most popular are histograms and Gaussian Mixtures.

Having defined skin color one can compute features from the amount and/or the spatial distribution of skin-colored pixels. It has been shown that these features may yield classifiers with reasonable performances. However, these approaches usually suffer from high false positive rates since not all images featuring large skin areas are necessarily adult images. Additionally not all skin-colored pixels actually depict human skin. Another drawback is that grayscale images cannot be classified using color features. Since the shape-based approaches, however, rely on color-information as well, they face the same problem.

There are also many approaches which use shape-based features for describing skin-colored regions of interest or their respective contours. Also, color-based features or moment invariants are often computed for such regions of interest as features. Other approaches use MPEG-7 descriptors based on shape as well as on color and texture as feature vectors.

The third group of approaches is based on local feature descriptors such as SIFT and Self-Similarity. These approaches use a visual dictionary to discretize the respective feature space. That way, the occurrence of local feature prototypes can be counted resulting in a visual word-occurrence vector for each image. The occurrence

vectors can then be used directly as image representations or mapped into a topic space of a pLSA model yielding a low-dimensional topic vector for each image.

For classifying feature vectors several machine-learning algorithms have been employed. The most popular one is the *k*-nearest neighbor classifier. However, SVM and MLP have also been used frequently.

Overall, approaches using color and shape information appear to be more robust than pure color-based approaches, since on average they produce less false positives at comparable or superior true positive rates. Of course, these observations are based on the numbers published by the respective authors which have been computed on different datasets. It is still reasonable that augmenting the color-based approaches increases the classification performance. The local feature-based approaches arguable perform at least on par with the remaining approaches as far as they can be compared.

7.2 Results and conclusion

Table 4 provides an overview of all papers on visual adult image recognition discussed in this survey (except for papers which present identical approaches). The number of test images is given as well as the ratio of true positives and false positives for each approach. However, this overview only summarizes the results and is not meant for comparison of the approaches. A direct comparison of the performance numbers reported in the various papers is not possible, since there is no common adult dataset for obvious reasons.

This is a major problem when comparing different approaches, since besides featuring different contents, the databases used by different researches greatly vary in size. Some experiments are conducted on barely more than one hundred images, while others use thousands or tens of thousands of images.

Note also that the difficulties of the test sets greatly vary among the different approaches. For instance, in [20] swim suit images are used and for the results stated in Table 4 they are counted as false positives if classified as adult. Another example is the negative test set of Bosson et al. [4] which explicitly contains a subset of images showing people. Presumably some of the other approaches do not include such difficult negative images. If the difficulty of the dataset is increased on purpose as for example by Desealers et al. [9] the classification performance declines significantly. It stands to reason that similar observations could be made for most of the other approaches.

Due to the lack of comparability, progress at the task of adult image recognition is seemingly achieved considerably slower than in other fields, since it is hard to figure out which among the known practices one should choose to integrate into new approaches. If researchers could agree on a few negative datasets and a surrogate positive set for adult images (e.g. people in swimwear), one could at least compare performance statistics.

Another important problem is that all approaches presented in this survey except the one by Ulges and Stahl [33] only consider legal adult images whereas images which are relevant under criminal law aspects in most countries (e.g. involving children) might significantly vary regarding image contents. This assumption is clearly confirmed by the experimental results of Ulges and Stahl. However, adult image filter systems should obviously also detect such images.

Table 4 Summary of the approaches described in this survey (where numbers were available)

References	# pos	# neg	% TP	% FP	Remarks
Duan et al. [10]	312	710	93.91	29.86	Color-based filter only.
Duan et al. [10]	312	710	80.70	10.00	
Jones and Rehg [18]	5453	5226	85.80	7.50	
Zheng et al. [42]	1297	3787	80.00	8.00	Approx. TP and FP taken from ROC curve.
Zheng et al. [42]	1331	50629	77.00	20.00	Approx. TP and FP taken from ROC curve on data of [31].
Rowley et al. [31]	1331	50629	84.00	20.00	Approx. TP and FP taken from ROC curve.
Hu et al. [16]	1000	1000	92.80	6.00	
Arentz and Olstad [1]	500	800	92.10	10.60	
Zheng et al. [44]	180	146	89.20	15.30	5-fold cross validation.
Ka [19]	2000	240	93.60	7.00	Subjectively ambiguous images did not affect the classification rates.
Wang et al. [35]	437	10809	97.50	18.40	
Wang et al. [35]	437	10809	92.05	10.07	Different thresholds.
Zeng et al. [41]	11349	59057	76.50	5.00	
Yoo [40]	2400	400	99.25	23.00	1,300 positive training images.
Yoo [40]	2400	2000	90.96	5.75	10,000 positive training images.
Kim et al. [21]	2694	2703	94.69	4.90	
Kim et al. [20]	<1700	<1700	91.37	7.35	
Bosson et al. [4]	3967	7038	83.87	10.86	SVM
Bosson et al. [4]	3967	7038	84.57	11.57	MLP
Fleck et al. [11]	138	1401	87.30	7.40	Color-based filter only.
Fleck et al. [11]	138	1401	52.20	3.40	
Lopes et al. [25]	<180	<180	84.60	na	5-fold cross validation. Ratio might be accuracy instead of TP.
Deselaers et al. [9]	~5100	~400	95.30	3.40	5-fold cross validation.
Deselaers et al. [9]	1000	1000	78.80	9.40	5-fold cross validation. Different dataset.
Ulges and Stahl. [33]	1000	1000	90.30	9.70	5-fold cross validation.
Ulges and Stahl. [33]	1000	1000	94.00	6.00	5-fold cross validation. Different negative set.
Lienhart and Hauke [24]	2068	12511	92.70	1.90	

The true positives (TP) and false positives (FP) are given for each approach as well as the respective number of adult (# pos) and non-adult (# neg) test images. The horizontal lines indicate in which section (Sections 4, 5, or 6) the respective approaches are described. Please note that this table is a mere summary and not intended for comparison of the approaches

We also want to point out that the runtime of adult image filter systems is a crucial issue. As mentioned in the introduction of this survey, one major application is filtering images on the Internet which means that approaches must be applicable in web-scale scenarios and preferably run in real time. Unfortunately, runtime evaluation and scalability is often omitted in research publications and thus we cannot compare the various approaches with regards to runtime. Intuitively, however, color-based approaches are in general less time-consuming than shape-based or local feature-based approaches. For instance, Google apparently uses (among other filtering techniques) the approach by Jones and Rehg [18] who use efficient color-based features. On the other hand, recent feature descriptors like histograms

of oriented gradients (HOG) [8] or speeded up robust features (SURF) [2] can also be computed rapidly.

In conclusion the problem of visual adult image recognition is a difficult task which is also reflected by the diversity of the different approaches described in this survey. Yet for the same reason it is a highly interesting topic and the importance of building reliable recognition systems will grow in the near future. Duan's résumé in his work of 2002 is still valid today: "Searches capable of distinguishing clearly among nudes, marmalades and national flags are still an unrealized dream" [10].

References

1. Arentz WA, Olstad B (2004) Classifying offensive sites based on image content. Comput Vis Image Underst 94:295–310
2. Bay H, Ess A, Tuytelaars T, Gool LV (2008) Surf: speeded up robust features. Comput Vis Image Underst (CVIU) 110(3):346–359
3. Bober M (2001) Mpeg-7 visual shape descriptors. IEEE Trans Circuits Syst Video Technol 11(6):716–719
4. Bosson A, Cawley G, Chan Y, Harvey R (2002) Non-retrieval: blocking pornographic images. In: Proceedings of the international conference on image and video retrieval, pp 50–60
5. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8(6):679–698
6. Choi B, Chung B, Ryou J (2009) Adult image detection using bayesian decision rule weighted by svm probability. In: Proceedings of the 2009 4th international conference on computer sciences and convergence information technology, ICCIT '09, pp 659–662
7. Chong CW, Raveendran P, Mukundan R (2003) A comparative analysis of algorithms for fast computation of zernike moments. Pattern Recogn 36(3):731–742
8. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the IEEE conference Computer Vision and Pattern Recognition 2005. CVPR 2005., pp 1–4
9. Deselaers T, Pimenidis L, Ney H (2008) Bag-of-visual-words models for adult image classification and filtering. In: Proceedings of the 19th International Conference on Pattern Recognition, 2008. ICPR 2008., pp 1–4
10. Duan L, Cui G, Gao W, Zhang H (2002) Adult image detection method base-on skin color model and support vector machine. In: Proceedings of the 5th Asian conference on computer vision, pp 797–800
11. Fleck MM, Forsyth DA, Bregler C (1996) Finding naked people. In: Proceedings of the European conference on computer vision, vol 2, pp 592–602
12. Gershon R, Jepson AD, Tsotos JK (1986) Ambient illumination and the determination of material changes. J Opt Soc Am **3**(10), 1700–1707
13. Gomez G, Morales EF (2002) Automatic feature construction and a simple rule induction algorithm for skin detection. In: Proceedings of the ICML workshop on machine learning in computer vision, pp 31–38
14. Hammami M, Chahir Y, Chen L (2006) Webguard: a web filtering engine combining textual, structural and visual content-based analysis. IEEE Trans Knowl Data Eng 18(2):272–284
15. Hu MK (1962) Visual pattern recognition by moment invariants. IRE Trans Inf Theory 8(2): 179–187
16. Hu W, Wu O, Chen Z, Fu Z, Maybank S (2007) Recognition of pornographic web pages by classifying texts and images. IEEE Trans Pattern Anal Mach Intell 29(6):1019–1034
17. Jebara TS, Pentland A (1997) Parametrized structure from motion for 3d adaptive feedback tracking of faces. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 144–150
18. Jones MJ, Rehg JM (2002) Statistical color models with application to skin detection. Int J Comput Vis 46(1):81–96
19. Ka CH (2009) A study on adult image detection via object analysis and multiresizing. In: Proceedings of the 9th international symposium on communications and information technology, 2009. ISCIT 2009, pp 784–789

20. Kim W, Lee HK, Park J, Yoon K (2005) Multi class adult image classification using neural networks. In: Advances in artificial intelligence, vol 3501. Springer, Berlin/Heidelberg, pp 222–226
21. Kim W, Lee HK, Yoo S, Baik S (2005) Neural network based adult image classification. In: Artificial neural networks: biological inspirations ICANN 2005, vol 3696. Springer, Berlin/Heidelberg, pp 481–486
22. Kovač J, Peer P, Solina F (2003) Human skin colour clustering for face detection. In: Proceedings of the IEEE Region 8 EUROCON 2003. Computer as a tool, vol 2, pp 144–148
23. Liao WH, Liu MJ (2004) Robust swimming style classification from color video. In: Proceedings of the interntaional computer symposium, pp 541–546
24. Lienhart R, Hauke R (2009) Filtering adult image content with topic models. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 2009. ICME 2009, pp 1472–1475
25. Lopes APB, de Avila SEF, Peixoto ANA, Oliveira RS, de A. Araújo A (2009) A bag-of-features approach based on hue-sift descriptor for nude detection. In: Proceedings of the 17th European Signal Processing Conference. EUSIPCO 2009, pp 1552–1556
26. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
27. Manjunath B, Ohm JR, Vasudevan V, Yamada A (2001) Color and texture descriptors. IEEE Trans Circuits Syst Video Technol 11(6):703–715
28. McKenna SJ, Gong S, Raja Y (1998) Modelling facial colour and identity with gaussian mixtures. Pattern Recogn 31(12):1883–1892
29. Ries CX, Lienhart R (2010) Automatic pose initialization of swimmers in videos. In: Proceedings of SPIE-IS&T electric imaging: visual information processing and communication, vol 7543, pp 75,430J–1–75,430J–8
30. Ries CX, Romberg S, Lienhart R (2010) Towards universal visual vocabularies. In: Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME), pp 1067–1072
31. Rowley HA, Jing Y, Baluja S (2006) Large scale image-based adult-content filtering. In: Proceedings of the 1st international conference on computer vision theory, pp 290–296
32. Shechtman E, Irani M (2007) Matching local self-similarities across images and videos. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2007. CVPR '07, pp 1–8
33. Ulges A, Stahl A (2011) Automatic detection of child pornography using color visual words. In: Proceedings of the 2011 IEEE International Conference on Multimedia and Expo. ICME 2011, pp 1–6
34. Vezhnevets V, Sazonov V, Andreeva A (2003) A survey on pixel-based skin color detection techniques. In: Proceedings of the GraphiCon-2003, pp 85–92
35. Wang JZ, Wiederhold G, Firschein O (1997) System for screening objectionable images using daubechies' wavelets and color histograms. In: Proceedings of the 4th international workshop on Interactive Distributed Multimedia Systems and telecommunication services, IDMS '97, pp 20–30
36. Wu O, Zuo H, Hu W, Zhu M, Li S (2008) Recognizing and filtering web images based on peoples existence. In: Proceedings of the IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08, pp 648–654
37. Yang J, Fu Z, Tan T, Hu W (2004) A novel approach to detecting adult images. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, vol 4, pp 479–482
38. Yang J, Lu W, Waibel A (1998) Skin-color modeling and adaptation. In: Proceedings of the Asian conference on computer vision, vol 2, pp 687–694
39. Yang MH, Ahuja N (1999) Gaussian mixture model for human skin color and its applications in image and video databases. In: Proceedings of SPIE '99, pp 458–466
40. Yoo SJ (2004) Intelligent multimedia information retrieval for identifying and rating adult images. Knowledge-Based Intelligent Information and Engineering Systems, pp 164–170
41. Zeng W, Gao W, Zhang T, Liu Y (2004) Image guarder: an intelligent detector for adult images. In: Proceedings of the Asian conference on computer vision 2004, pp 198–203
42. Zheng H, Daoudi M, Jedynak B (2004) Blocking adult images based on statistical skin detection. Electron Lett Comput Vis Image Anal 4(2):1–14
43. Zheng H, Liu H, Daoudi M (2004) Blocking objectionable images: adult images and harmful symbols. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 2004. ICME '04, vol 2, pp 1223–1226
44. Zheng QF, Zeng W, Wen G, Wang WQ (2004) Shape-based adult images detection. In: Proceedings of the 3rd international conference on image and graphics, 2004, pp 150–153

Christian X. Ries is a PhD student at the Multimedia Computing Lab of the University of Augsburg. He acquired the Master degree in Computer Science from the University of Augsburg in 2009. His research interests are in the areas of Computer Vision, Machine Learning and Image Content Analysis.

Rainer Lienhart is a full professor in the computer science department of the University of Augsburg. He received his Ph.D. in Computer Science from the University of Mannheim, Germany, in 1998, where he was a member of the Movie Content Analysis Project (MoCA). From August 1998 to July 2004 he was a Staff Researcher at Intel's Microprocessor Research Lab in Santa Clara, California, where he worked on transforming a network of heterogeneous, distributed computing platforms into an array of audio/video sensors and actuators capable of performing complex DSP tasks such as distributed beamforming, audio rendering, audio/visual tracking and camera array processing. At the same time, he was also continuing his work on media mining, where he is well-known for his work in video content analysis with contributions in text detection/recognition, commercial detection, face detection, shot and scene detection and automatic video abstraction.

He has been a committee member of ACM Multimedia, IEEE International Conference on Multimedia Systems, IEEE International Conference on Multimedia Expo, SPIE Storage and Retrieval of Media Databases, IEEE Workshop on Content-Based Access of Image and Video libraries and the International Eurographics Workshop on Multimedia. He is a reviewer for IEEE Transactions on Pattern Recognition and Machine Learning, IEEE Transactions on Multimedia, Journal of Comput Vision and Image Understanding and ACM Multimedia Systems Journal.

Dr. Lienhart has published over 50 papers in major conferences and journals and filed 20+ patents.