Similarity of Color Images

Markus Stricker and Markus Orengo

Communications Technology Laboratory

Swiss Federal Institute of Technology, ETH

CH-8092 Zurich, Switzerland

stricker@vision.ee.ethz.ch

Abstract

We describe two new color indexing techniques. The first one is a more robust version of the commonly used color histogram indexing. In the index we store the cumulative color histograms. The L_1-, L_2-, or L_∞-distance between two cumulative color histograms can be used to define a similarity measure of these two color distributions. We show that while this method produces only slightly better results than color histogram methods, it is more robust with respect to the quantization parameter of the histograms. The second technique is an example of a new approach to color indexing. Instead of storing the complete color distributions, the index contains only their dominant features. We implement this approach by storing the first three moments of each color channel of an image in the index, *i.e.*, for a HSV image we store only 9 floating point numbers per image. The similarity function which is used for the retrieval is a weighted sum of the absolute differences between corresponding moments. Our tests clearly demonstrate that a retrieval based on this technique produces better results and runs faster than the histogram-based methods.

Keywords: color indexing, color distribution features, cumulative histograms.

1   Introduction

The paradigm of color indexing into an image database works as follows: Given a query image, we want to retrieve all the images whose color compositions are similar to the color composition of the query image. Color indexing is based on the observation that often color is used to encode functionality: Road signs are painted with a specific set of colors, roads are black, forests are green, sky is blue etc. An object's color will not allow us to determine its identity. In general, texture or

even geometric properties are needed to identify objects. Consequently, color indexing methods are bound to retrieve false positives, *i.e.*, images with completely different content which just happen to have a similar color composition as the query image. Thus, in practice it is necessary to combine color indexing with texture and/or shape indexing methods (see [Niblack *et al.* 1993, Pentland *et al.* 1994, Gong *et al.* 1994]). Even if texture and shape indexing methods improve, we believe that color indexing will retain its importance as a computationally simple and fast filter whose output is then processed by computationally more intensive methods. In this context the major challenge of new color indexing methods is to improve the robustness of finding images with similar color compositions and to increase the retrieval speed.

Color indexing was introduced by [Swain and Ballard 1991]. They store coarsely quantized color histograms of the images in the index. Under some assumptions their histogram similarity measure corresponds to the metric which is induced from the L_1-norm on the histogram space. Changes in lighting and changes due to occlusion may cause relatively large changes in their similarity measure. [Funt and Finlayson 1991] improved the robustness with respect to changes in the lighting by histogramming color ratios. Yet, their tests show that their color constant color indexing performs only marginally better than Swain and Ballard's method. [Niblack *et al.* 1993] propose to determine the similarity of color histograms with an L_2-related metric. The application of their metric produces less false negatives and it produces a different and not necessarily smaller set of false positives than the application of the L_1 metric. This suggests that the use of color histograms is the major source of instability.

In this paper we present two new color indexing methods. The first one was stimulated by the attempt to make the above mentioned methods more robust. The index contains the complete color distributions of the images in the form of cumulative color histograms. The color distributions are compared using the L_1-, the L_2-, or the L_∞-metric. The second method is an example of a completely different approach to color indexing. Instead of storing the complete color distributions, we store only their major features. A good choice of the features can increase the robustness of the retrieval. In the retrieval process only these features are compared and thus, it is significantly faster than a retrieval process based on comparing complete color distributions.

In the next section we explain the weaknesses of the existing methods. In section 3 we show how color distributions can be stored in the index such that the similarity measures based on the L-metrics become more robust. The new concept of only working with color distribution features is introduced in section 4. In the same section we give an intuitive set of features which yield an index of minimal size and which robustly support a computationally fast similarity measure. In section 5 we compare the results of our methods with the results obtained by applying the L_1- and the L_2-metric to color histogram indices. We conclude the article with a remark on how our methods can be improved.

2 Shortcomings of existing methods

In this section we discuss the shortcomings of the most frequently used similarity measures on the color histogram space: The L_1-metric [Swain and Ballard 1991, Funt and Finlayson 1991] and L_2-related metrics [Niblack *et al.* 1993].

If we map the colors in the image M into a discrete color space containing n colors, then the color histogram $H(M)$ is a vector $(h_{c_1}, h_{c_2}, \ldots, h_{c_n})$, where each element h_{c_j} represents the number of pixels of color c_j in the image M. Without loss of generality we may assume that all images contain N pixels and hence $\sum_{i=1}^{n} h_{c_i} = N$.

The L_1-distance of two color histograms H and I is defined as

$$d_{L_1}(H, I) = \sum_{l=1}^{n} |h_{c_l} - i_{c_l}| \ .$$

We have already noticed in [Stricker 1994] that a retrieval which is based on the L_1-distance produces many false negatives, *i.e.*, it does not retrieve all the images with perceptually similar color histograms. This happens because perceptually similar color histograms may be a large L_1 distance apart from each other. Changes in lighting which may result in a slight shift in the color histogram cause the L_1-metric to misjudge the similarity completely. If we order the bins of the color histograms in a way such that neighboring bins correspond to similar colors, then the distance between the histograms H_1 and H_2 in figure 1 should certainly be smaller than the one
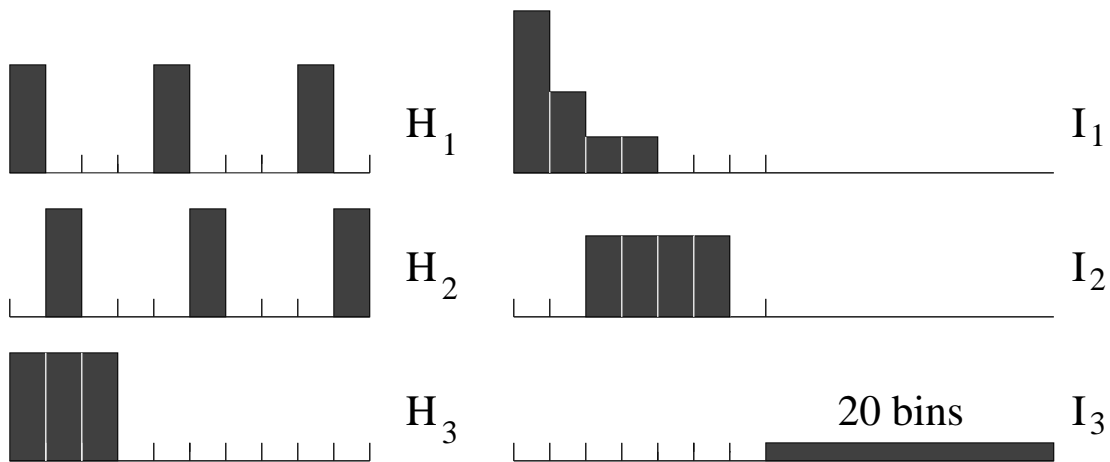

Figure 1: Six color histograms whose preceptual similarities do not correspond to their L_1- and their L_2-distances.

between H_1 and H_3. But the distances are

$$d_{L_1}(H_1, H_2) = 2N > d_{L_1}(H_1, H_3) = d_{L_1}(H_2, H_3) = 1.33N \ . \tag{1}$$

This clearly shows the undesired effect of using the L_1-distance as a similarity measure of color histograms. In large the problem can be attributed to the fact that the L_1-metric does not take the color similarity between the bins into account.

The metric defined by [Niblack *et al.* 1993] makes use of the color similarities of the bins in the color histograms. If the color similarity of the j-th and the l-th bin is a_{jl} and the symmetric matrix A with entries a_{jl} is positive definite, then it defines a metric

$$d_A(H, I) = \sqrt{(H - I) \cdot A \cdot (H - I)^T}$$

on the color histogram space. With a basis transform A can be diagonalized. Thus, there exists a discrete color space in which the different colors have no correlation at all. If we denote the colors of this new color space by d_l, then

$$d_A(H, I) = \sqrt{\sum_{l=1}^{n} w_l \cdot (h_{d_l} - i_{d_l})^2} \, ,$$

where w_l are the eigenvalues of the matrix A. Hence, the idea of working with a color similarity matrix simply reduces to a weighted L_2-metric in a suitably chosen color space. We discuss only the case in which all weights are 1, *i.e.* $d_A = d_{L_2}$, but the following can also be demonstrated for any other set of weights. To a less extreme degree, d_{L_2} suffers from the same problem as d_{L_1}: For the color histograms displayed in figure 1 the L_2 distance are

$$d_{L_2}(H_1, H_2) = 0.82N > d_{L_2}(H_1, H_3) = d_{L_2}(H_2, H_3) = 0.66N \, . \tag{2}$$

The large number of false positives in an L_2-based retrieval can mostly be attributed to the fact that the L_2 distance overestimates the similarity of color distributions without a pronounced mode, *i.e.*, with many non-empty bins. We demonstrate this with the three color histograms I_1, I_2 and I_3 which are displayed in figure 1. Their L_2 distances are

$$d_{L_2}(I_1, I_2) = 0.68N > d_{L_2}(I_1, I_3) = 0.63N > d_{L_2}(I_2, I_3) = 0.55N \tag{3}$$

which does not correspond to the natural order of similarity.

3   Similarity of color distributions

So far research has focused on finding new distance functions which correspond better to the perceptual similarity of color histograms. But even incorporating color similarities into the distance function does not yield a robust distance function which corresponds to the preceptual similarity of color histograms. The information which is stored in the index and the algorithm that is used to retrieve images with similar color compositions are closely related components of color indexing. Thus, we believe that color indexing can only be improved substantially if the data which is stored

in the index contains a more robust characterization of the color distributions *and* a more efficient retrieval algorithm can be designed.

The color composition of an image can be viewed as a color distribution in the sense of probability theory. Notice that the discrete form of a probability distribution is a cumulative histogram. These distributions can be of almost any shape. Methods from non-parametric statistics can be used to determine whether two arbitrarily shaped, empirical distributions represent the same distribution. Most of the non-parametric statistical tests work only for one dimensional distributions and thus, they cannot be applied to three dimensional color distributions. Fortunately, we do not need a rigorous statistical test, *i.e.*, we do not need to know the probability with which two empirical distributions represent the same distribution. A function which yields a relative order of color distributions suffices for color indexing purposes. In other words, we only need a rank order of the color distributions in the index with respect to similarity with the color distribution of the query image. The discrete, three dimensional analogues of the Kolmogorow-Smirnow test (L_∞-metric), of the Cramèr - von Mises test (L_2-metric), or the L_1-metric are appropriate for this purpose [Siegel 1956, Breiman 1973].

To describe the index we first have to fix an order of the colors in the discrete color space. Normally we work with the natural order: The color $c_j = (r_j, g_j, b_j)$ is less than the color $c_l = (r_l, g_l, b_l)$ if $r_j < r_l$, $g_j < g_l$, and $b_j < b_l$. The cumulative color histogram $\tilde{H}(M) = (\tilde{h}_{c_1}, \tilde{h}_{c_2}, \ldots, \tilde{h}_{c_n})$ of the image M is defined in terms of the color histogram $H(M)$:

$$\tilde{h}_{c_j} = \sum_{c_l \leq c_j} h_{c_l} \ .$$

To imitate the above mentioned non-parametric statistical tests we have to store the cumulative color histograms of the images in the index. To determine the similarity of these histograms we use one of the following three measures:

$$d_{L_1}(\tilde{H}, \tilde{I}) = \sum_{j=1}^n |\tilde{h}_{c_j} - \tilde{i}_{c_j}| \quad , \quad d_{L_2}(\tilde{H}, \tilde{I}) = \sqrt{\sum_{j=1}^n (\tilde{h}_{c_j} - \tilde{i}_{c_j})^2} \quad , \quad d_{L_\infty}(\tilde{H}, \tilde{I}) = \max_{1 \leq j \leq n} |\tilde{h}_{c_j} - \tilde{i}_{c_j}| \ .$$

The cumulative histograms of the histograms in figure 1 are shown in figure 2. The distances between the cumulative histograms are

$$d_{L_1}(\tilde{H}_2, \tilde{H}_3) = 4N > d_{L_1}(\tilde{H}_1, \tilde{H}_3) = 3N > d_{L_1}(\tilde{H}_1, \tilde{H}_2) = N$$
$$d_{L_2}(\tilde{H}_2, \tilde{H}_3) = 1.41N > d_{L_2}(\tilde{H}_1, \tilde{H}_3) = 1.20N > d_{L_2}(\tilde{H}_1, \tilde{H}_2) = 0.58N$$
$$d_{L_\infty}(\tilde{H}_2, \tilde{H}_3) = d_{L_\infty}(\tilde{H}_1, \tilde{H}_3) = 0.67N > d_{L_\infty}(\tilde{H}_1, \tilde{H}_2) = 0.33N$$

and

$$d_{L_1}(\tilde{I}_1, \tilde{I}_3) = 15.63N > d_{L_1}(\tilde{I}_2, \tilde{I}_3) = 13N > d_{L_1}(\tilde{I}_1, \tilde{I}_2) = 2.63N$$
$$d_{L_2}(\tilde{I}_1, \tilde{I}_3) = 3.43N > d_{L_2}(\tilde{I}_2, \tilde{I}_3) = 3.01N > d_{L_2}(\tilde{I}_1, \tilde{I}_2) = 1.23N$$
$$d_{L_\infty}(\tilde{I}_1, \tilde{I}_3) = d_{L_\infty}(\tilde{I}_2, \tilde{I}_3) = N > d_{L_\infty}(\tilde{I}_1, \tilde{I}_2) = 0.75N \ .$$
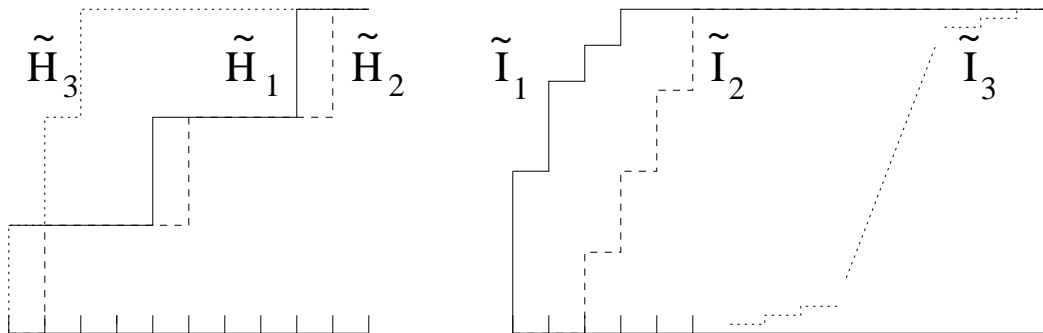
Figure 2: The cumulative histograms of the histograms displayed in figure 1.

The order resulting from the L_1- and the L_2-distance correspond to the intuitive order of the similarities of the color histograms. The orders resulting from the L_∞-distance contain an equality where we would prefer similarity values that are close to each other. Notice that these orders differ from the orders obtained from the L_1-distance (1) and the L_2-distance (2) and (3) of the corresponding histograms. In section 5 we will show that cumulative color histograms together with the L-metrics are more robust with respect to the quantization of the histograms than the L-distances applied to color histograms.

The cumulative color histograms are always completely dense vectors even if only a few colors of the discrete color space appear in each image. Usually, the non-empty bins in color histograms are sparse and thus, the price we pay for the increased robustness of our new method is an increase in the index size and a slower retrieval speed. Since the robustness of these techniques allows us to work with very coarsely quantized color spaces, the increase in storage requirements and computational resources are only marginal.

4   Similarity of color distribution features

In this section we present a new approach to color indexing and we describe one possibility of how this approach can be implemented.

The quantization of the bins in the color histograms and in the cumulative color histograms is a parameter which is necessary because of the method that we apply to solve the indexing problem, *i.e.*, it is not intrinsically related to the problem itself. This makes it hard to find an optimal quantization. Furthermore, even an optimal quantization will produce the usual, unwanted quantization effects. The search for a method without parameters in the index creation process lead to the approach of working with color distribution features.

The idea of using color distribution features for color indexing is simple. In the index we store dominant features of the color distributions. It is imperative that these features can be extracted robustly from the images. The retrieval process is based on a similarity function which uses only these features to determine the similarity of color distributions. This approach has the potential to outperform the histogram-based methods in the robustness of the results as well as in the retrieval speed.

From probability theory we know that a probability distribution is uniquely characterized by its moments, resp. central moments. Thus, if we interpret the color distribution of an image as a probability distribution, then the color distribution can be characterized by its moments, as well. We propose to store the first moment, and the second and the third central moment of each color channel in the index. The first moment is the average and hence, we store the average color of the image. The second and the third central moment are the variance and the skewness of each color channel. We store the standard deviation and the third root of the skewness of each color channel in the index. Hence, all the values in the index have the same units which makes them somewhat comparable. If the value of the i-th color channel at the j-th image pixel is p_{ij}, then the index entries related to this color channel are:

$$E_i = \frac{1}{N} \sum_{j=1}^{N} p_{ij} \;, \quad \sigma_i = \left(\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^2\right)^{\frac{1}{2}} \;\; \text{and} \;\; s_i = \left(\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^3\right)^{\frac{1}{3}} \;.$$

Let H and I be the color distributions of two images with r color channels. If the index entries of these images are E_i resp. F_i, σ_i resp. ς_i, and s_i reps. t_i, then we define the similarity as

$$d_{\mathrm{mom}}(H, I) = \sum_{i=1}^{r} w_{i1}|E_i - F_i| + w_{i2}|\sigma_i - \varsigma_i| + w_{i3}|s_i - t_i| \;,$$

where $w_{kl} \geq 0$ ($1 \leq l, k \leq 3$) are user specified weights. Since we work only with a small subset of the moments of the color distributions, the function d_{mom} is not a metric, *i.e.*, it is possible that two non-identical color distributions have a similarity value of 0. This is the reason why we call d_{mom} a similarity function and not a similarity measure or a metric. A retrieval based on d_{mom} may produce false positives because the index contains no information about the correlation between the color channels. Nevertheless, we will demonstrate in section 5 that this method is more robust than the other methods that we discussed.

The weights in d_{mom} can be used to tune the similarity function for a given application. For example, if we know that all the images in the database were taken under the same lighting conditions, then we set $w_{i1} > w_{i2}$ and $w_{i1} > w_{i3}$ in order to penalize shifts in the average color. If the database contains images of outdoor scenes, then it is very likely that the lighting conditions differ for all the images. In this case, the average color contains unreliable information which can be expressed in the similarity function by setting w_{i1} to small values. Very often we work with images in the HSV color space and we want the hue to match more strictly than the saturation and the value. This can be achieved by setting all the weights for the moments of the hue channel to a higher value than the other weights.

If we set all the weights to 1, then we can qualitatively estimate the moments of the color distributions in figure 1. Their similarity values would certainly be

$$d_{\mathrm{mom}}(H_2, H_3) > d_{\mathrm{mom}}(H_1, H_3) > d_{\mathrm{mom}}(H_1, H_2)$$

and

$$d_{\mathrm{mom}}(I_1, I_3) > d_{\mathrm{mom}}(I_2, I_3) > d_{\mathrm{mom}}(I_1, I_2) \,,$$

which is the correct order of these values.

5   Discussion of Results

We use an image database that contains 3000 color jpeg images. They are 8-bit color images of size 185×123. They cover a wide range of nature scenes, buildings, construction sites, animals, etc. To decouple the color channels at least partially, we perform all the tests in the HSV color space. We build 2 different indices to evaluate the robustness of the histogram related methods. In the first index 16 bins cover the range of the hue and 4 bins cover the range of the saturation, respectively the value. In the second index the histograms are quantized much coarser: We have 8 bins for the hue and 2 bins for the saturation, as well as 2 bins for the value. We present the results for the similarity function that works with the moments of the color distributions for the three weight matrices given in table 1. Notice that since the range of the hue is between 0 and 2π,

Table 1: Three weight matrices for comparing the moments of color distribution.

	H	S	V			H	S	V			H	S	V
average	1	2	1		average	1	2	3		average	1	2	1
std. dev.	1	2	1		std. dev.	1	2	3		std. dev.	2	4	2
skewness	1	2	1		skewness	1	2	3		skewness	2	4	2
	W_1					W_2					W_3		

and the ranges of the saturation and the value are between 0 and 1, the weight matrices given in table 1 emphasize the matching of the hue. We chose query images for which there exist absolutely obvious matches in the database. For each query image we report its rank in the sorted list of similarity values. A good color indexing method assigns low ranks to the obvious matches. We have run a large number of tests and the results that we present below are representative for the performance of these methods.

The first query image shows two gas tanks. The database contains 3 obvious matches which are images of the same scene from different viewing angles and with a slightly different background. The ranks for these images are displayed in table 2. The second query image shows an owl on top

Table 2: Ranks of the obvious matches of two gas tanks in a database of size 3000.

query image:

index sim. measure		rank of the image			max. rank
9 moments	W_1	4	5	8	8
	W_2	2	8	6	8
	W_3	4	6	9	9
8/2/2	L_∞	34	98	79	98
16/4/4	L_∞	3	57	42	57
cum. hist. 8/2/2	L_1	53	162	30	162
16/4/4	L_1	33	354	8	354
8/2/2	L_2	65	158	34	158
16/4/4	L_2	15	306	11	306
8/2/2	L_1	138	394	48	394
16/4/4	L_1	4	132	6	132
histogram 8/2/2	L_2	71	541	102	541
16/4/4	L_2	10	1358	75	1358

of a tree trunk. The database contains 6 obvious matches which are images of the same animal after it has moved a little and in which different portions of its shadow and the tree trunk appear. The changes in the ranks due to different choices of the parameters are displayed in table 3.

The indexing method based on the comparison of the moments is clearly superior to the other methods. We have experimented with 10 different weighting matrices. For the gas tank query image the worst rank of an obvious match is 13 and for the owl query image it is 28. Thus, this method is very robust and even the worst case performance is better than any of the histogram based techniques. For some other query images, the advantage of working with the moments was even more pronounced: The worst rank of an obvious match for the moment-based method was 10, while the worst ranks for the histogram-based techniques with different quantizations were all on the order of 1000. In the majority of the queries the L-metrics applied to the cumulative color histograms yield slightly better results than the same metrics applied to color histograms. Table 3 displays the results of an exceptional case in which L-metrics applied to color histograms produced better results than the same metrics applied to cumulative color histograms. The main advantage of working with the cumulative histograms is that the L-metrics become more stable with respect to the quantization parameter. Our tests indicate that the winner among the L-metrics depends on the query image. Furthermore, the L-metrics do not always yield better results if they are applied to histograms or cumulative histograms with a higher quantization. The latter two remarks clearly show how difficult it is to tune a histogram-based technique optimally.

6   Conclusion

We have presented two new color indexing techniques. For the first method we have to store cumulative color histograms in the index and to determine the similarity of index entries we apply the L_∞-metric. On average this method does not yield better match values than the application of the L_1- or the L_2-metric to compare cumulative color histograms, but it is extremely robust with respect to the quantization parameter of the cumulative histograms. The second technique is an example of a new approach to color indexing. Instead of storing the complete color distribution we only store a few of its dominant features. We chose to work with the first three moments of the distribution in each color channel. We have constructed a simple similarity function which is based on these features of the color distributions. Our test results clearly show that this approach outperforms the other techniques: Its index is the smallest, the retrieval process is the fastest and it produces the best results.

Work is now in progress to further improve the second method by incorporating the correlation between the color channels and thus, treating a color distribution as a true three dimensional distribution rather than as three separate one dimensional distributions.

query image:

		rank of the image						max.
index	sim. measure							rank
9 moments	W_1	3	4	5	8	12	13	13
	W_2	2	1	3	7	10	16	16
	W_3	5	4	7	11	14	15	15
8/2/2	L_∞	1	2	3	40	61	26	61
16/4/4	L_∞	1	3	2	26	18	135	135
cum. hist 8/2/2	L_1	1	2	4	31	60	26	60
16/4/4	L_1	1	2	3	30	24	57	57
8/2/2	L_2	1	2	4	40	66	25	66
16/4/4	L_2	1	2	3	29	22	104	104
8/2/2	L_1	1	2	3	19	29	8	29
16/4/4	L_1	1	2	3	9	6	49	49
histogram 8/2/2	L_2	1	2	3	24	50	20	50
16/4/4	L_2	1	3	2	12	6	82	82

Table 3: Ranks of the obvious matches of an owl in a database of size 3000.

Acknowledgement

References

[Breiman 1973] L. Breiman. *Statistics: With a view towards applications*. Houghton Mifflin Company, 1973.

[Funt and Finlayson 1991] B. V. Funt and G. D. Finlayson. Color constant color indexing. Technical Report 91-09, School of Computing Science, Simon Fraser University, Vancouver, B.C., Canada, 1991.

[Gong *et al.* 1994] Y. Gong, H. Zhang, et al. An image database system with content capturing and fast image indexing abilities. In *Proc. of the IEEE International Conference on Multimedia Computing and Systems*, pages 121–130, May 1994. Boston, Mass.

[Niblack *et al.* 1993] W. Niblack, R. Barber, et al. The QIBC project: Querying images by content using color, texture and shape. In *Storage and Retrieval for Image and Video Databases I*, volume 1908 of *SPIE Proceedings Series*, Feb. 1993.

[Pentland *et al.* 1994] A. Pentland, R. W. Picard, S. Sclaroff, et al. Photobook: Tools for content-based manipulation of image databases. In *Storage and Retrieval for Image and Video Databases II*, volume 2185 of *SPIE Proceedings Series*, Feb. 1994.

[Siegel 1956] S. Siegel. *Nonparametric statistics: For the behavioral sciences*. McGraw-Hill, Inc., 1956.

[Stricker 1994] M. Stricker. Bounds for the discrimination power of color indexing techniques. In *Storage and Retrieval for Image and Video Databases II*, volume 2185 of *SPIE Proceedings Series*, pages 15–24, Feb. 1994.

[Swain and Ballard 1991] M. J. Swain and D. H. Ballard. Color indexing. *Intern. Journal of Computer Vision*, 7(1):11–32, 1991.