A Maximum Likelihood Investigation Into Color Indexing

Nicu Sebe Michael S. Lew
Leiden Institute of Advanced Computer Science,
Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands
{nicu mlew}@liacs.nl

Abstract

Color is an important attribute for image matching and retrieval. Most of the attention from the research literature has been focussed on the color model with little or no consideration of the noise models. In this paper we investigate the problem of color indexing from a maximum likelihood perspective. We take into account the color model, the noise distribution, and the quantization of the color features. Furthermore, from the real noise distribution we derive a distortion measure which consistently provides improved accuracy. Our investigation concludes with results on a real stock photography database consisting of 11,000 color images.

1 Introduction

As the world enters the digital age, visual media is becoming prevalent and easily accessible. Factors such as the explosive growth of the World Wide Web, terabyte disk servers, and the digital versatile disk, reveal the growing amount of visual media which is available to society. With the availability of visual media comes the associated problem of searching for it and consequently, the focus of researchers toward providing automatic content based retrieval systems. Of the visual media retrieval methods, color indexing is one of the dominant methods because it has been shown to be effective in both the academic and commercial arenas. In color indexing, histogram methods are often used because they are feasible in terms of memory usage and provide sufficient accuracy. The histogram methods quantize each image into a feature vector based on a color model such as RGB [4] or HSV [4], and then compare the query image feature vector to the database image feature vectors using a minimum distance classifier.

In previous works such as [3] and [10], comparisons have been made between different distance metrics. However, their results did not explain why a particular metric would provide better results. Here in our work, we show that the maximum likelihood paradigm explains why one metric will outperform another one based upon the underlying noise model. Furthermore, we show how to derive a better distortion measure based upon the real noise distribution.

1.1 Color Models

Color models describe different aspects of the color space of an image. Two frequently used models are RGB and HSV. RGB refers to the intensity of 3 additive color primaries, red, green, and blue. Each primary is typically quantized into 256 levels and then combined to create $256 \times 256 \times 256$ possible colors. The HSV model separates the color components from the luminance component. The hue and saturation of a color are represented by H and S, and the luminance is represented by V.

When we create a color histogram, we must quantize each component of the color model using a number of bits. We define quantization X:Y:Z for color model ABC as quantizing color component A using X bits, B using Y bits, and C using Z bits. In the case of HSV, a 4:2:2 quantization refers to quantizing H using 4 bits, S using 2 bits, and V using 2 bits. When not otherwise specified RGB refers to a 3:3:2 quantization and HSV refers to a 4:2:2 quantization.

We chose to use 11,000 images from the Corel Photo database because it represents a widely used set of photos by both amateur and professional graphical designers. Furthermore, it is available on the Web at http://www.corel.com.

1.2 Color Indexing

In color indexing, given a query image, the goal is 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: grass is green, sky is blue etc.

If we map the colors in the image $\mathcal{I}$ into a discrete color space containing n colors, then the color histogram [11, 9] $H(\mathcal{I})$ is a vector $(h_{c_1}, h_{c_2}, \cdots, h_{c_n})$, where each element h_{c_j} represents the probability of having the color c_j in the image $\mathcal{I}$.

Two widely used distance metrics are L_1 [5] and L_2

(a) (b) (c) (d)

Figure 1: Two examples of test copy pairs used (a)-(c) the original image; (b)-(d) copy image

[1]. For example the L_1 distance applied to two color histograms H and I is defined as

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

Similarly, the L_2 distance will be

$$d_{L_2}(H,I) = \sqrt{\sum_{i=1}^{n} (h_{c_i} - i_{c_i})^2}$$

1.3 Early Experiments

Before we can measure the accuracy of particular methods, we first had to find a challenging and objective ground truth for our tests. We perused the typical image alterations and categorized various kinds of noise with respect to finding image copies. Copies of images were often made with images at varying JPEG qualities, in different aspect ratio preserved scales, and in the printed media. We defined these as JPEG noise, Scaling noise, and Printer-Scanner noise. JPEG noise was created by coding and then decoding a JPEG image using varying JPEG-quality values. Using HSV 4:2:2 and JPEG quality 30, we were able to recover the exact image copy as the top rank with 100% accuracy from our large image database. In Scale noise, we made the copy by reducing the image in size so that the image was aspect ratio preserved with maximum size 32×32. Using HSV 4:2:2, the copy was found within the top 10 ranks with 100% accuracy. We concluded that JPEG noise and Scaling noise were not sufficiently challenging to separate the different color indexing methods.

In Printer-Scanner noise, the idea was to measure the effectiveness of a retrieval method when trying to find a copy of an image in a magazine or newspaper. We printed 110 images using an Epson Stylus 800 color printer at 720 dots per inch, and then scanned each of them using an HP IIci color scanner. These 110 copy pairs formed our ground truth test set. When comparing a query image to a database image, we normalized them to have the same mean in order to avoid gray-level bias. The noise from this copy process was the most significant in that the copy was found in the top 10 ranks using HSV 4:2:2 with less than 35 % accuracy. Examples of the test copy pairs are shown in Figure 1. In the copy pair containing the child, note that the textures on the sleeve and on the hair are missing. Also, the cup and water jug are barely discernible. In the other copy pair, note the loss of details in the background mountainside and windows on the lower-right house wall. Furthermore, note that we purposely chose a hard test set in order to have a good discrimination between the retrieval methods.

1.4 Usability Issues

In creating a system for users, it is important to take into account the way in which users will interact with the system. Two important issues are the total response time of the system and the number of results pages which the user must look at before finding the image copy. We make the following assumptions. First, in order to have an interactive experience, the total system response time should be less than 2 seconds. Furthermore, the number of results pages which are looked at by the user should reflect the usage of real professionals. Graphical artists typically flip through stock photo albums containing hundreds of pages, which amounts to a few thousand images for relevant material. For this reason we show the results regarding the top 1 to 6000 ranks. We also avoid methods which require more than a few seconds of response time.

Section 2 describes the mathematical support for maximum likelihood approach. In Section 3 we examine the modeling of the noise distribution along with the retrieval accuracy with respect to the color model and quantization. Furthermore, we investigate the usage of the maximum likelihood theory towards deriving an ideal distortion measure for the real noise distribution. Discussion and Conclusions are given in Section 4.

2 Maximum Likelihood Approach

From the mathematical-statistical point of view, the problem of finding the right model for the similarity noise comes down to the maximization of the similarity probability.

Consider first, two subsets of M images from the database (D) : $X \subset D, Y \subset D$ which according to the ground truth are similar:

$$X \equiv Y \tag{1}$$

This can be written:

$$x_i \equiv y_i, \qquad i = 1, ..., M \tag{2}$$

where x_i and y_i represent the feature vectors associated with the images in the corresponding subsets.

The equation (2) can be further written as:

$$x_i = y_i + n_i, \qquad i = 1, ..., M \tag{3}$$

where n_i represent the "noise" vector obtained as the difference between the two vectors.

In this context the similarity probability can be defined:

$$P(X, Y) = \prod_{i=1}^{M} \{\exp[-\rho(x_i, y_i)]\} \tag{4}$$

where function ρ is the negative logarithm of the probability density of the noise.

According to (4) we have to find the probability density function of the noise that maximizes the similarity probability: *maximum likelihood* estimate for the noise distribution [7].

We can further consider that this noise distribution is valid for all of the database, so using it for all of the images in the database one obtains the best possible ranking results.

Taking the logarithm of (4) we find that we have to minimize the expression:

$$\sum_{i=1}^{M} \rho(x_i, y_i) \tag{5}$$

In this case, according to (3), the function ρ depends only on the difference between its two arguments. We can replace (5) with:

$$\sum_{i=1}^{M} \rho(n_i) \tag{6}$$

where $n_i \equiv x_i$-y_i and the operation "-" denotes difference between corresponding values in the feature vectors.

To analyze the behavior of the estimate we take the approach described in [6] and [8] based on an *influence function*. The influence function characterizes the bias that a particular measurement has on the solution and is proportional to the derivative, ψ, of the estimate [2].

$$\psi(n) \equiv \frac{d\rho(n)}{dz} \tag{7}$$

In case the noise is Gaussian distributed:

$$Prob\{x_i - y_i\} \sim \exp(-[x_i - y_i]^2) \tag{8}$$

then

$$\rho(n) = n^2 \qquad \psi(n) = n \tag{9}$$

If the errors are distributed as a *double* or *two-sided exponential*, namely

$$Prob\{x_i - y_i\} \sim \exp(-|x_i - y_i|) \tag{10}$$

then, by contrast,

$$\rho(n) = |n| \qquad \psi(n) = sgn(n) \tag{11}$$

In this case, using (5) we minimize the *mean absolute deviation*, rather than the *mean square deviation*. Here the tails of the distribution, although exponentially decreasing, are asymptotically much larger than any corresponding Gaussian.

For normally distributed errors, (9) says that the more deviant the points, the greater the weight. By contrast, when tails are somewhat more prominent, as in (10), then (11) says that all deviant points get the same relative weight, with only the sign information used.

In summation, one can note that (8) resembles the L_2 metric while (10) resembles the L_1 metric. Thus, the *maximum likelihood* approach gives a direct connection between the noise distribution and the comparison metrics. If ρ is the negative logarithm of the probability density of the noise, then the corresponding metric is given by (5).

3 Experiments

As we stated in Section 1.3, JPEG noise and Scaling noise were not sufficiently challenging to separate the different color indexing methods therefore, we focused on Printer-Scanner noise application. Our ground truth consists of 110 copy-pairs: the original images along with their copies obtained by printing and then scanning.

3.1 Distribution Analysis

The first question we asked was, "Which distribution is the closest to the real color model noise?" To answer this we needed to measure the noise with respect to each color model and then we could choose the color model and noise which had the best accuracy.

According to (3) the real noise distribution is obtained as the normalized histogram of differences between the elements of color histograms corresponding to copy-pair images that form the ground truth.

In Figures 2 and 3 we display the real noise distribution in RGB and HSV respectively. Note that the best fit exponential has a better fit to the noise distribution than

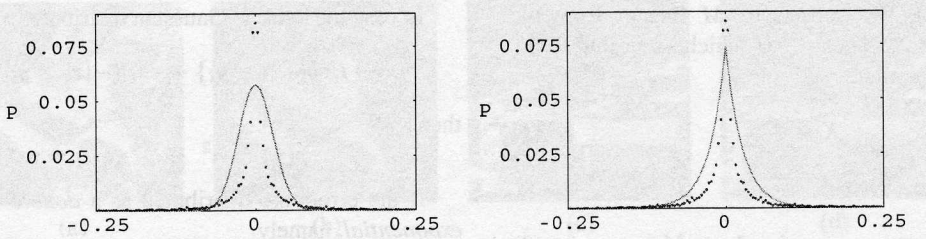


Figure 2: Similarity noise distribution in RGB compared to best fit Gaussian (a) (modeling error is 0.11) and best fit exponential (b) (modeling error is 0.085)

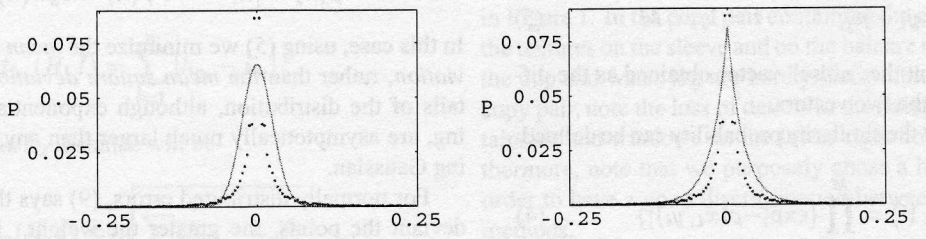


Figure 3: Similarity noise distribution in HSV compared to best fit Gaussian (a) (modeling error is 0.106) and best fit exponential (b) (modeling error is 0.082)

the Gaussian for both color models. Consequently, this implies that the L_1 metric will give better retrieval accuracy than the L_2 in both cases. For the retrieval accuracy we choose to display percentage of correct copies found within the top n matches. From the tests as shown in Figure 5 it is clear that the L_1 metric gives a significant improvement in retrieval accuracy as compared to L_2.

3.2 Color Model

The second question we asked was, "Which color model gives better retrieval accuracy?". As shown in Figure 4 using the L_1 metric we obtained an improvement in retrieval accuracy by up to 8% when using the HSV color model.

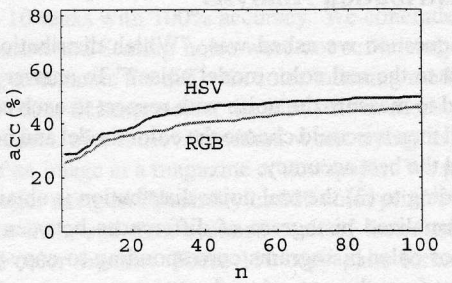


Figure 4: Retrieval accuracy for the top 100 using L_1 - RGB vs. HSV

3.3 Quantization

Based upon the improvement in the retrieval accuracy it is clear that the best choice is to use the HSV color model with the L_1 metric. So, the next question is, "How does the quantization scheme affect the retrieval accuracy?". In Figure 6(a) it appears that increased resolution in H may be the cause of increased accuracy. This leads us to ask whether further H resolution will give even better results. Figure 6(b) shows that this is not the case.

In summary, the experiments in this section showed that the choice of color model, noise distribution, and quantization can affect the accuracy by up to 8%, 15%, and 7%, respectively.

3.4 Ideal Distribution

If it is necessary to perform analytic computations, then the usage of one of the analytic metrics like, L_1 or L_2 metrics is required. The main advantage of these metrics is the ease in implementation and analytic manipulation. However, neither distance measure models the real noise distribution accurately, so we expect that we can lower the misdetection rates further. Using equations (4)-(6), we extract from the test database noise distribution a distortion measure within the maximum likelihood paradigm, which we denote as the ML distortion measure. The ML distortion measure is directly related to the real noise distribution which is a discrete distribution with known points. Con-

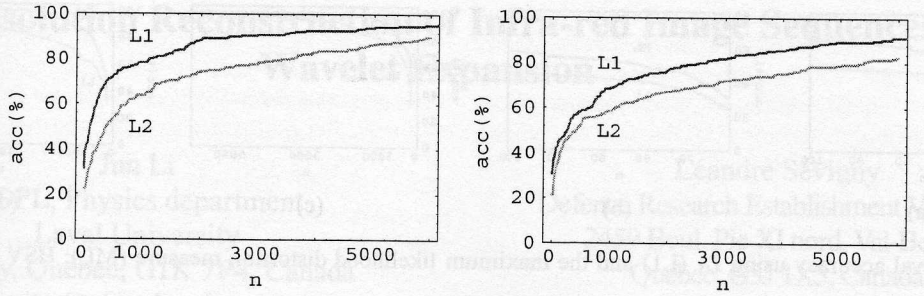


Figure 5: Retrieval accuracy for the top 6000 matches (a) HSV (b) RGB

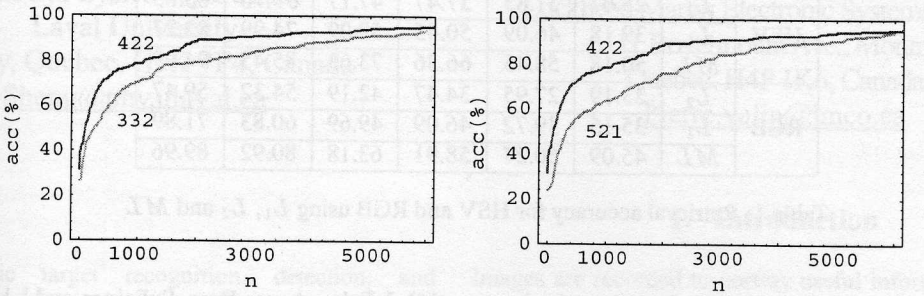


Figure 6: Retrieval accuracy for HSV using different quantization models (a) 4:2:2 - 3:3:2 and (b) 4:2:2 - 5:2:1

sider that we have to compare two vectors (histograms) then, for each difference value between corresponding elements we have to calculate according to (5) the negative logarithm of the probability density of the real noise in that point. Since the distribution is discrete, the value of the probability in any arbitrary point is calculated by using interpolation between the two known adjacent probability values. The sum of all values calculated in this way resembles the ML distortion measure.

Since the L_1 measure outperformed the other measures in the previous sections, we displayed in Figure 7 the retrieval accuracy using the L_1 and ML distortion measures. Note that the ML distortion measure consistently has better retrieval accuracy. Table 1 summarizes the results for retrieval accuracy for L_1, L_2 and ML.

In summary, regarding a new and effective method for color indexing, we have presented the theory of maximum likelihood in Section 2, evaluated commonly used metrics, and created an optimized distortion measure based on the real noise distribution which gives significantly improved results over the commonly used metrics.

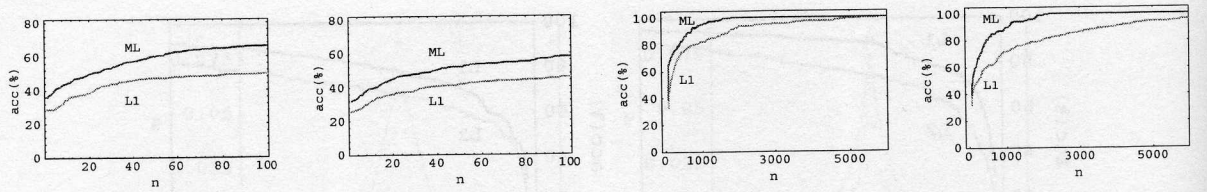
4 Discussion and Conclusions

In this paper we investigated the problem of color indexing for content based retrieval using the maximum likelihood paradigm. The maximum likelihood theory provides us with a direct connection between the noise distribution and the retrieval accuracy of the system. We tested the maximum likelihood based methods on an 11,000 stock image database and found the following results

- HSV beats RGB.
- L_1 beats L_2.
- H carries more information than S.
- ML beats L_1 by significant margins.
- Color distributions are not Gaussian.

Note that we deliberately chose a hard test set. The numerical results we obtained reflect this. We were also concern about the relevance of the user needs: some users may be interested in the improved accuracy in the top 100, while other users, like graphical artists will be interested in a global improved accuracy across the entire database. Therefore, it is important to have an improved accuracy even for top 20 or more ranks.

This paper presents maximum likelihood as a unifying theory for color indexing algorithms. Previous work has identified empirical facts such as the L_1 metric gives better accuracy, but none of the past research has given a detailed theoretical justification for the improvement. The



(a) (b) (c) (d)

Figure 7: Retrieval accuracy using L_1 (L1) and the maximum likelihood distortion measure (ML): HSV (a)-(c), RGB (b)-(d)

Top		20	40	100	200	500	1000
HSV	L_2	25.67	31.83	37.47	47.17	59.76	65.83
	L_1	39.18	46.09	50.45	59.09	74.99	82.27
	ML	50.18	58.45	66.36	73.63	85.45	94.99
RGB	L_2	23.19	27.95	34.47	42.19	54.32	59.47
	L_1	35.15	39.72	46.09	49.69	60.83	71.89
	ML	45.09	50.27	58.91	63.18	80.92	89.96

Table 1: Retrieval accuracy for HSV and RGB using L_1, L_2 and ML

first point of this paper has been to show how the color indexing algorithms are special cases of the maximum likelihood approach as applied to specific noise distributions.

Second, maximum likelihood theory clearly describes the breaking points of an algorithm. Given a representative sample, the noise distribution can be estimated, and then maximum likelihood theory can be directly used to determine the efficacy of a particular metric.

Third, we have shown that significant accuracy improvement can be achieved by using an ideal distortion measure based on the real noise distribution. Maximum likelihood theory provides both the framework and the method for deriving the ideal distortion measure.

References

[1] A. Berman and L.G. Sapiro. Efficient image retrieval with multiple distance measures. *Proc. SPIE, Storage and Retrieval for Image/Video Databases*, 3022:12–21, 1997.

[2] M.J. Black. *Robust Incremental Optical Flow*. PhD thesis, Yale University, September 1992.

[3] M. Flicker, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[4] J. Foley, A. van Dam, S. Feiner, and J. Hughes. *Computer Graphics - principles and practice*. Addison-Wesley, 1990.

[5] A. Gupta, S. Santini, and R. Jain. In search of information in visual media. *Communications ACM*, 12:34–42, 1997.

[6] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistic: The Approach Based on Influence Functions*. John Wiley and Sons, New York, 1986.

[7] P.J. Huber. *Robust Statistic*. NewYork: Wiley, 1981.

[8] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons, New York, 1987.

[9] H.S. Sawhney and J.L. Hafner. Efficient color histogram indexing. In *Proc. of 1994 IEEE International Conference on Image Processing*, volume 2, pages 66–70, 1994.

[10] J.R. Smith. *Integrated Spatial and Feature Image Systems: Retrieval, Compression and Analysis*. PhD thesis, Columbia University, February 1997.

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