

# Robust Color Indexing

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

## ABSTRACT

In content based image retrieval, color indexing is one of the most prevalent retrieval methods. In literature, most of the attention 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

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 [2] or HSV [2], and then compare the query image feature vector to the database image feature vectors using a minimum distance classifier.

In previous works such as [7] and [9], comparisons have been made between different distance metrics. However, their results did not explain why a particular metric would provide better results. Here 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 Indexing

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: grass is green, sky is blue, etc.

If we map the colors in the image  $Q$  into discrete color space containing  $n$  colors, then the color histogram [10, 8]  $H(Q)$  is a vector  $(h_{c_1}, h_{c_2}, \dots, h_{c_n})$ , where each element  $h_{c_j}$  represents the number of pixels of color  $c_j$  in the image  $Q$ .

Two widely used distance metrics are  $L_1$  [3] and  $L_2$  [1]. Other criterion functions that have been used in previous literature are (1) histogram intersection [10], which is equivalent with  $L_1$ , (2) average color distance [4], (3) the quadratic distance measure form [6].

### 1.2 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, for an interactive environment, 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.

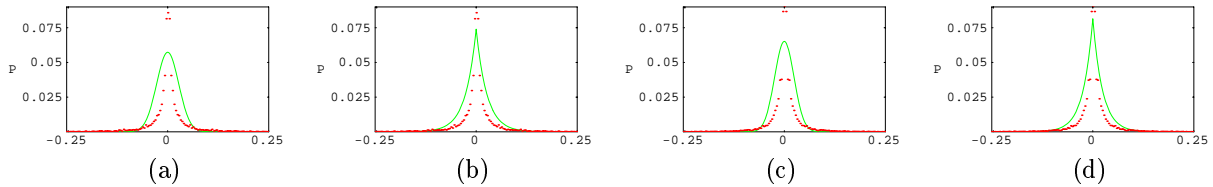


Figure 1: Similarity noise distribution in RGB (a),(b) compared to best fit Gaussian (a) (modeling error is 0.15) and best fit exponential (b) (modeling error is 0.09); Similarity noise distribution in HSV (c),(d) compared to best fit Gaussian (c) (modeling error is 0.106) and best fit exponential (d) (modeling error is 0.082);

## 2 MAXIMUM LIKELIHOOD ESTIMATOR

Consider two subsets of  $M$  images from the database ( $D$ ):  $X \subset D$ ,  $Y \subset D$  which according to the ground truth are similar. Let  $x_i$  and  $y_i$ , with  $i = 1, \dots, M$ , be the feature vectors associated with the images in the corresponding subsets ( $X$  and  $Y$ , respectively) and let the noise  $n_i$  be the distortion between  $x_i$  and  $y_i$ . In this context, we can define the similarity probability as follows:

$$P(X, Y) = \prod_{i=1}^M \{\exp[-\rho(n_i)]\} \quad (1)$$

where function  $\rho$  is the negative logarithm of the probability density of the noise.

According to (1) we have to find the probability density function of the noise that maximizes the similarity probability: *maximum likelihood* estimate for the noise distribution [5].

Due to space limitations we restrict to some considerations. Maximum likelihood gives a direct connection between the noise distribution and the comparison metrics. Using the maximum likelihood theory, one can easily prove that when the noise distribution is Gaussian, the corresponding metric is  $L_2$ . In this case, the maximum likelihood estimate is obtained by minimizing the *mean square deviation*. If the noise is distributed as a double or two-sized exponential, the maximum likelihood estimate is obtained by minimizing the *mean absolute deviation* and therefore, the corresponding metric is  $L_1$ . For a general noise distribution, considering  $\rho$  as the negative logarithm of the probability density of the noise, the corresponding metric is given by equation (2).

$$\sum_{i=1}^M \rho(n_i) \quad (2)$$

## 3 EXPERIMENTS

In our experiments, 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>.

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. The first two alterations were not sufficiently challenging since the copy was found within the top 10 ranks with 100% accuracy. 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. Note that we purposely chose a hard test set in order to have a good discrimination between the retrieval methods.

### 3.1 Distribution Analysis, Color Model and Quantization

The first question we asked was, "Which distribution is a good approximation for 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.

The real noise distribution is obtained as the normalized histogram of differences between the elements of color histograms corresponding to copy-pair images from ground truth.

In Figure 1 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

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 2, it is clear that the  $L_1$  metric gives a significant improvement in retrieval accuracy as compared to  $L_2$ .

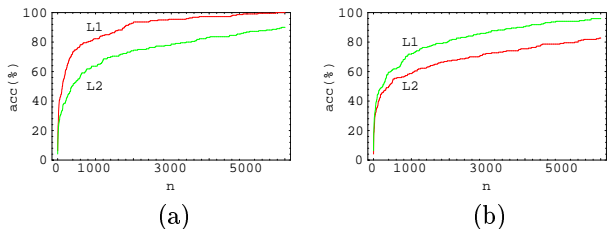


Figure 2: Retrieval accuracy for the top 6000 matches (a) HSV (b) RGB

The second question we asked was, "Which color model gives better retrieval accuracy?". We considered the RGB and HSV color spaces, and using the  $L_1$  metric we obtained an improvement in retrieval accuracy by up to 8% when using the HSV color model.

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?". We considered different quantization schemes for HSV color space and we found that the best choice for our application is HSV 4:2:2. Note that a 4:2:2 quantization refers to quantizing H using 4 bits, S using 2 bits, and V using 2 bits.

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 5%, respectively.

### 3.2 Ideal Distribution

If it is necessary to perform analytic computations, then the usage of one of the analytic metrics like  $L_1$  or  $L_2$ , 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 the real noise distribution we extract a distortion measure within the maximum likelihood paradigm, which we denote as the  $ML$  distortion measure. This measure is directly related to the real noise distribution which is a discrete distribution with known points. Consider that we have

to compare two vectors (histograms), then, for each difference value between corresponding elements we have to calculate according to Eq. (2) 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 3 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 briefly 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

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  $L_2$ .
- $ML$  beats  $L_1$  by significant margins.
- Color distributions are not Gaussian.

Note that we deliberately chose a hard test set and the numerical results we obtained reflect this. We were also concerned 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.

## 5 CONCLUSIONS

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

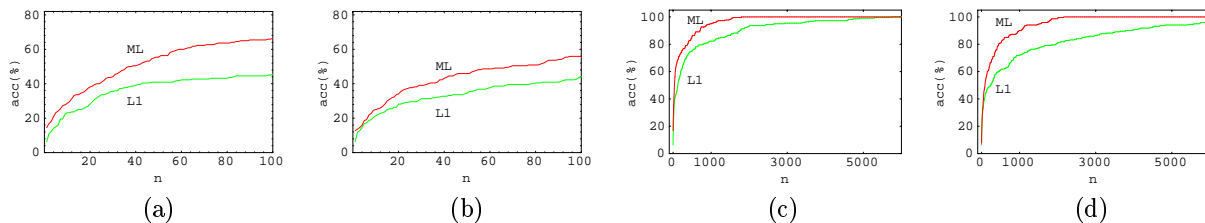


Figure 3: Retrieval accuracy using  $L_1$  (L1) and the  $ML$  distortion measure (ML): HSV (a)-(c), RGB (b)-(d)

Top		20	40	100	200	500	1000
HSV	$L_2$	23.15	28.17	36.42	42.17	51.76	65.83
	$L_1$	28.18	39.09	45.45	59.09	74.99	82.27
	$ML$	38.18	50.45	66.36	73.63	85.45	94.99
RGB	$L_2$	19.17	24.81	32.15	38.19	46.32	59.47
	$L_1$	24.15	32.72	41.09	49.69	60.9	71.89
	$ML$	34.09	43.18	55.9	63.18	80.9	89.96

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

a detailed theoretical justification for the improvement. The 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.

## 6 ACKNOWLEDGEMENTS

This research was supported with a grant from Philips in the Netherlands.

## 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] J. Foley, A. van Dam, S. Feiner, and J. Hughes. *Computer Graphics-principles and practice*. Addison-Wesley, 1990.
- [3] A. Gupta, S. Santini, and R. Jain. In search of information in visual media. *Communic. ACM*, 12:34–42, 1997.
- [4] J. Hafner. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):729–736, 1997.
- [5] P.J. Huber. *Robust Statistic*. NewYork: Wiley, 1981.
- [6] W. Y. Ma, Y. Deng, and B.S. Manjunath. Tools for texture/color based search of images. *Proc. SPIE, Human Vision and electronic imaging II*, 3106:496–507, 1997.
- [7] W. Niblack, R. Barber, W. Equitz, M. Flicker, E. Glasman, D. Petrovic, P. Yanker, C. Faloutsos, and G. Yaublin. The QBIC project: Querying images by content using color, texture and shape. *SPIE - Storage and Retrieval for Image and Video Databases*, 1908:173–181, 1993.
- [8] 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.
- [9] J.R. Smith. *Integrated Spatial and Feature Image Systems: Retrieval, Compression and Analysis*. PhD thesis, Columbia University, February 1997.
- [10] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.