# Binary Histogram in Image Classification for Retrieval Purposes

Iivari Kunttu[1], Leena Lepistö[1], Juhani Rauhamaa[2], and Ari Visa[1]

[1]Tampere University of Technology
Institute of Signal Processing
P. O. Box 553, FIN-33101 Tampere, Finland
+358 (0)3 3115 4393
Iivari.Kunttu@tut.fi

[2]ABB Oy
Paper, Printing, Metals & Minerals
P. O. Box 94, FIN-00381 Helsinki, Finland
+358 (0)10 22 22630
Juhani.Rauhamaa@fi.abb.com

## ABSTRACT

Image retrieval can be considered as a classification problem. Classification is usually based on some image features. In the feature extraction image segmentation is commonly used. In this paper we introduce a new feature for image classification for retrieval purposes. This feature is based on the gray level histogram of the image. The feature is called binary histogram and it can be used for image classification without segmentation. Binary histogram can be used for image retrieval as such by using similarity calculation. Another approach is to extract some features from it. In both cases indexing and retrieval do not require much computational time. We test the similarity measurement and the feature-based retrieval by making classification experiments. The proposed features are tested using a set of paper defect images, which are acquired from an industrial imaging application.

## Keywords

Histogram, Indexing, Content-based image retrieval, Paper defect images

## 1. INTRODUCTION

The growth of digital imaging during the last few years has affected many fields of human life. Nowadays digital imaging is popularly used in many industrial solutions concerning e.g. quality control and process control. In the process industry several on-line measurement systems are based on the digital imaging. As a result of this development, the amount of the image data has increased rapidly. Consequently the sizes of different kinds of image databases in the industry have increased significantly. Therefore managing of these databases has become necessary.

The goal in the industrial imaging applications in many cases is to divide different images into different classes. Therefore a fast image classification system is necessary. The methods developed on the field of content-based image retrieval [Sme00] can be applied to this kind of image classification.



***Journal of WSCG, Vol.11, No.1., ISSN 1213-6972***
*WSCG'2003, February 3-7, 2003, Plzen, Czech Republic.*


In the retrieval process the images are retrieved from the database based on some features that characterize the image. In other words, image database indexing has to be done by extracting certain features from the images. In the existing content-based image retrieval systems the most common features are shape, color, and texture. Image segmentation [Gon93] is a necessary step if object specific features are to be used. This feature-based image selection is similar to common classification processes. Therefore in this paper, we consider the image retrieval as a classification problem.

In our research work we are concentrating on the retrieval of gray level paper defect images. The images are stored in large databases, which may contain tens of thousands of images. The objects in the images are paper defects and their background is flawless paper with small variation of gray levels. These defects can be for example holes, wrinkles, paper scraps or dirt. Until now the analysis of paper defects has based on the image segmentation [Iiv96], which is relatively time consuming process. Therefore there is a need for a method that is able to classify the defects without segmentation. Gray scale or color histograms are commonly used in image retrieval, in [Haf95], for example. Swain and Ballard

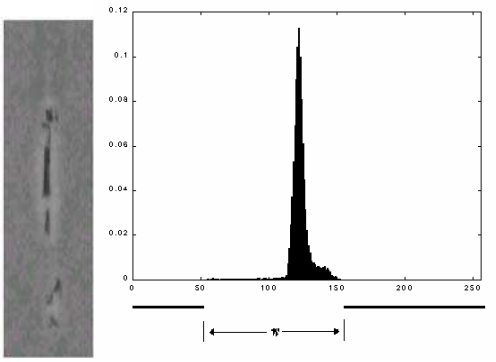


**Figure 1. Paper defect image, its 256-bin histogram and binary form of the histogram.**

[Swa90] introduced the original idea of using color histograms in indexing of large image databases. After that, color histograms have also been used in recognition of objects in the images [Enn95]. In this paper, we propose a method to utilize the image gray level histogram, a binary histogram. This binary histogram is a binary vector-form representation of the gray level distribution of the image. The main advantage of this approach is that it does not require any kind of image segmentation.

The principle of binary histograms is presented in section two. In the same section we introduce some image features that can be calculated based on the binary histogram. Another way to use binary histograms in retrieval is to define the similarity between them. In the end of the section two we present a similarity measure for the histograms. In section three, we test the effectiveness of retrieval ability of the proposed features. The purpose of this experiment is to find out, how well the features can distinguish between paper defect classes. This is done using a simple classifier.

## 2. BINARY HISTOGRAMS IN IMAGE CLASSIFICATION

The distributions of colors or gray levels of images are commonly used characteristics in the image analysis and classification. The gray level distribution can be presented as a gray level histogram:

$$H(G) = \frac{n_G}{n} \quad G = 0,1,...,NG-1 \qquad (1)$$

where $n_G$ is the number of pixels having gray level $G$, $n$ is the total number of pixels and $NG$ is the total number of gray levels.

### Binary Histogram

Image histogram and features calculated from it are not necessarily optimal for image classification without segmentation. Reason for that is the effect of the image background. The background causes a high peak to appear in the image histogram, whereas the defects occur as relatively small values in the histogram. Therefore, when the histograms of different images are compared e.g. by calculating distances between them, the peak caused by the background dominates the resulting distance. Consequently, the small values of the histogram, which are the most important ones, do not have significant effect on the comparison. For that reason, histogram itself does not work very well as a classifying feature for the paper defect images. Therefore there is a need for a solution that ignores or minimizes the effect of the image background.

Because the object (defect) in the image is represented by small values in the histogram, the bins representing these values should be extracted from the histogram in some way. One solution is to consider each bin of the histogram equally, which means that each bin of the histogram has the same significance, which is independent on the bin value. This can be made by quantizing the values in the histogram into two values only: "zero" and "nonzero". As a result, we obtain a binary vector $\mathbf{H_B}$, which is called binary histogram. The length of the vector is the same as the total number of the gray levels in the image ($NG$).

After the binary histogram has been calculated for the image, it can be used in two ways in classification. The first of the methods calculates some features which characterize the binary form of the histogram. In order to calculate these features, we have to define $\mathbf{G_B}$ as a set of gray levels which correspond to "1" in $\mathbf{H_B}$. Another classification method is based on the calculation of similarities between binary histograms

### Features of binary histogram

In this paper we use two types of features to characterize the gray level distribution of the image. The first of these features is the location of the gray level set $\mathbf{G_B}$ in the gray level scale of the image. The location depends on the gray level distribution of the object, and therefore it is essential feature in the paper defect classification. In addition to the location, also the size of $\mathbf{G_B}$ is important in classification. The size describes the variety of the gray levels represented by the image. In this section we present simple statistical means to measure these properties from the binary histograms.

The location of the distribution can be measured in many ways. One way to estimate it is to calculate the gravity center of $\mathbf{G_B}$ [Bal82]. In binary case, gravity center is equal to mean value ($mean$):

$$mean = mean(\mathbf{G_B}) \qquad (2)$$

The size of the distribution can be estimated using standard deviation (*std*):

$$std = \sqrt{\sum_{i=1}^{n} (\mathbf{G_B}(i) - mean)^2 / n} \quad (3)$$

in which $n$ is the amount of gray levels in the image. Another approach to the size estimation of the binary histogram is to calculate the width ($w$) of the distribution. This can be done by calculating the distance between smallest and largest components in the vector $\mathbf{G_B}$ (see figure 1):

$$w = \max(\mathbf{G_B}) - \min(\mathbf{G_B}) \quad (4)$$

## Similarity measurement between the binary histograms

In addition to the statistical measures which were presented in the previous section, there is also another way to utilize the binary histogram in image classification and retrieval. Because there is visible similarity between the binary histograms of similar defect images, the images belonging to the same class can be found by seeking similar binary histograms among the images. The binary histogram vectors can be matched by calculating similarity or distance measure between the vectors. Several different types of distance measures have been developed for similarity measurement [Dud01],[Han01]. Specific distance measures have also been introduced for binary data. These measures are based on the number of same and different elements in the binary vectors. When comparing two binary vectors of the same length, $\mathbf{H_{B1}}$ and $\mathbf{H_{B2}}$, let $n_{1,1}$ denote the number of the elements, whose value is in both vectors 1. In a similar way, $n_{1,0}$, $n_{0,1}$ and $n_{0,0}$ denote numbers of vector elements, which have values 1 and 0, 0 and 1, 0 and 0, respectively. *Jaccard coefficient* [Han01], is a popular measure for binary data. This coefficient is defined as:

$$S = \frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}} \quad (5)$$

When we consider the effectiveness of the image retrieval in terms of similarity measurement, it is obvious that the length of the binary vector plays a role. There are two reasons for that: 1) When we use this binary vector in indexing of large image database, it is clear that short vector is preferable. 2) When we retrieve the images from the database using similarity measures, the computing time depends strongly on the length of the binary vector. For these reasons the amount of the gray levels in the image histograms is very important aspect in the retrieval. We will investigate the effect of the vector length on the computing time and classification accuracy in section 3.

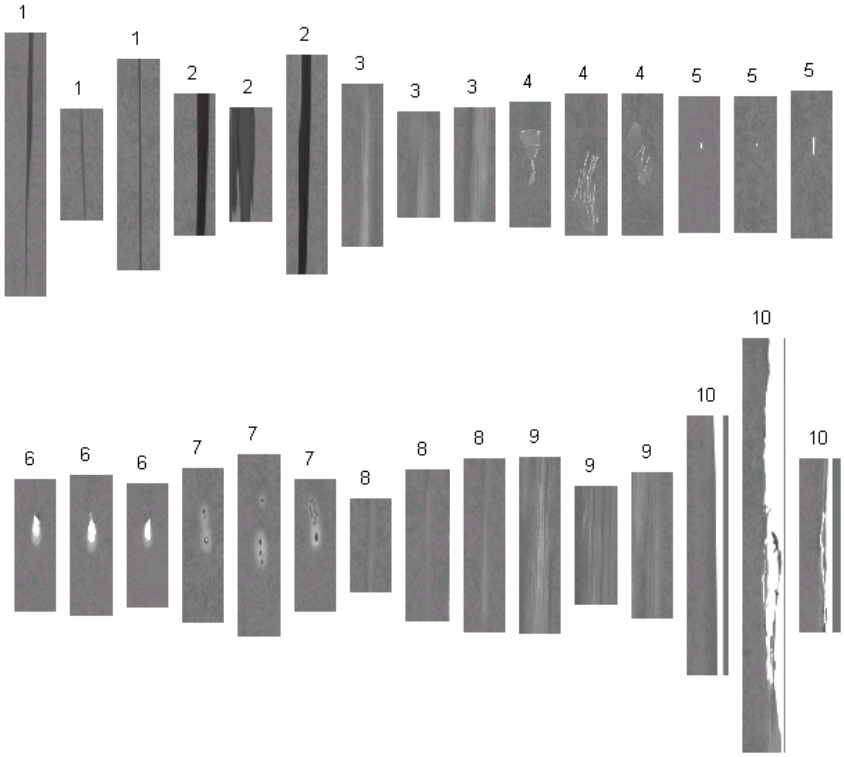


**Figure 2. Example images of paper defects. The numbers above images indicate the class of each defect.**

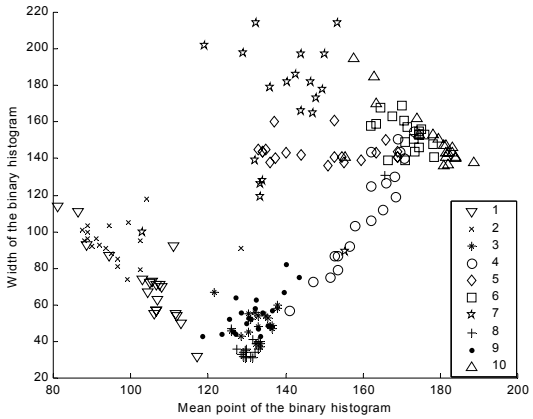


**Figure 3. Mean-width plot of the test set images.**

## 3. EXPERIMENTS

In the experimental part of our work we used a set of real paper defect images. The images were taken of paper web by a paper inspection system [Rau02]. The objects in the images are typical paper surface defects. The test set consisted of 200 paper defects, which represented ten defect classes with 20 defect images each. The images have 256 gray levels and their length varied strongly (from 100 to 2000 pixels). In figure 2, three example images from each class are presented.

The goal of the experimental part of this work was to clarify, how well the features and methods introduced in section 2 can distinguish between the defect image classes. The histogram was calculated for each image and after that the histograms were transformed into binary form. The features presented in section 2 were calculated for each histogram and the similarity measures between each histogram were defined as well. The calculation was made using Matlab on a PC with 804 MHz Pentium III CPU and 256 MB primary memory.

In the experimental part of this work the image retrieval task was considered as a classification problem. We tested the classifying ability of the binary histogram using a simple *k*-nearest neighbor classifier [Dud01]. The same classification principle was used in both feature-based and similarity-based classification. The principle of the *k*-nearest neighbor classifier is the following: The classifier selects *k* images from the test set that are closest to the query image. A class, to which most of those *k* images belong, is voted to be the class of the query image. In the cross-validation we used *leaving-one-out* method [Han01]. In this method each image is left out from the set in turn, and it is classified by means of the other images in the test set.

| Number of gray levels | Result of classification (%) | Computing time (sec) |
|---|---|---|
| 8 | 51.5 | 14 |
| 16 | 53 | 22 |
| 32 | 69 | 42 |
| 64 | 65.5 | 83 |
| 128 | 61 | 164 |
| 256 | 61 | 335 |

**Table 1. Results of classification using similarity measure**

### Classification based on the features

The features presented in section 2, mean point (*mean*), standard deviation (*std*) and the width (*w*) were selected to experimental part of this work. According to the experiments, the best classification results were obtained using feature combinations: mean-standard deviation and mean-distribution width. In these cases the features were able to classify 62.5% and 69% of the images into correct classes, respectively. The value of *k* in the classifier was three in both cases and computing time for classification of all 200 images was about 1.1 seconds. In the figure 3 is presented the median-width plot of the test set images. Figure 4 shows the mean result in each class.

### Classification based on the similarity measure

The effectiveness of the similarity measurement was tested using the Jaccard coefficient. The testing was made using the test set and 3-nearest neighbor classification.

In this experiment we decreased the amount of gray levels in the test set images. In this way we could investigate the effect of the binary histogram length on the computing time in classification. The results of the classification as well as the computing times of the whole classification process of 200 images are presented in table 1. The results show that the computing time decreases significantly when we use shorter binary vector. The length of the vector have effect also on the classification accuracy, and the optimal number of gray levels seems to be 32. In this case we obtain the best classification result (69% of the images are classified into correct classes) and the computing time has decreased 85% from the original. The classification results in each class using 32 gray levels are presented in figure 4.

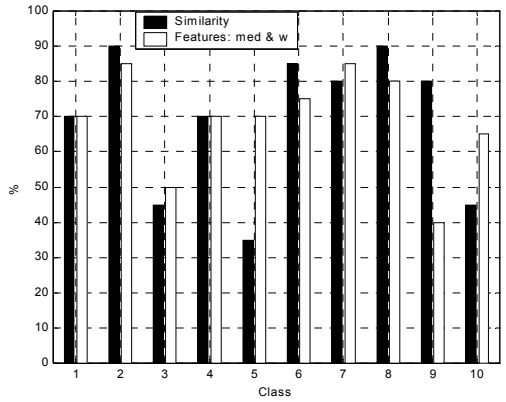


**Figure 4. The mean result of the classification in each class using similarity measure and features of figure 3.**

## 4. DISCUSSION

In this paper we presented a new approach to the histogram-based retrieval of the images. The method is a simple solution to the segmentation problem: Because we minimize the effect of the peak caused by the image background in the histogram, we do not need any time-consuming segmentation, which makes the method effective in many cases.

For testing purposes we had a set of paper defect images, whose classification is a quite demanding task. The results of the classification experiments showed that the binary histogram is a suitable feature for retrieval of those images. However, there are differences in the results of different defect classes (figure 4), and even better results probably could be achieved by using additional features in retrieval. These additional features could characterize for example the shape and size of paper defects. On the other hand, the classification was correct in 69% of the images using only binary histogram. This result is significant, because visible differences between certain defect classes are very small.

In section two we presented two different approaches to the indexing of the paper defect images. Features calculated by means of the binary histogram proved to be powerful in terms of classification accuracy and computing time. Another approach, similarity measurement based on Jaccard coefficient, gave similar result in accuracy, but it required more computing time. On the other hand, this similarity-based approach is more general way to make image indexing, because the features can be calculated from binary vectors also during retrieval. Consequently the both approaches seem to be useful in the image retrieval. In addition, according to our experiments, decrease of the image gray levels or image resolution do not affect significantly on the classification accuracy.

Based on the results presented in this paper, it is obvious that the binary histogram is able to classify paper surface defects. However, the principles presented here can be generalized to other similar image classification and retrieval problems. Methods based on the binary histograms offer a simple and fast way to make image indexing for retrieval purposes without segmentation.

## 5. ACKNOWLEDGMENTS

The authors wish to thank the Technology Development Center of Finland (TEKES's grant 40397/01) for financial support.

## 6. REFERENCES

[Bal82] Ballard, D. H., and Brown, C. M. Computer Vision, Prentice Hall, New Jersey, 1982

[Dud01] Duda, R. O., Hart, P. E., and Stork, D. G. Pattern Classification, 2nd edition, John Wiley & Sons, 2001.

[Enn95] Ennesser, F., and Medioni G. Finding Waldo, or Focus of Attention Using Local Color Information, In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 17, No 8, pp. 805-809, Aug 1995.

[Gon93] Gonzalez, R. C., and Woods, R. E. Digital Image Processing, Addison Wesley, 1993.

[Haf95] Hafner, J., Sawhney, H. S., Equitz, W., Flickner, M., and Niblack, W. Efficient Color Histogram Indexing for Quadratic Form Distance Function, In IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 17, No 7, pp. 729-739, July 1995.

[Han01] Hand, D., Mannila, H., and Smyth, P. Principles of Data Mining, MIT Press, Massachusetts, 2001.

[Iiv96] Iivarinen, J., Rauhamaa, J., and Visa, A. Unsupervised segmentation of surface defects, Proceedings of 13th International Conference on Pattern Recognition, Wien, Austria, Vol. 4, pp. 356-360, Aug. 25–30, 1996.

[Rau02] J. Rauhamaa, R. Reinius: Paper Web Imaging with Advanced Defect Classification, Proceedings of the 2002 TAPPI Technology Summit, Atlanta, Georgia, March 3-7, 2002.

[Sme00] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-Based Image Retrieval at the End of the Early Years, In IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 22, No 12, pp. 1349-1379, December 2000.

[Swa90] Swain, M. J., and Ballard, D. H. Indexing Via Color Histograms, In Proceedings of third international conference on Computer Vision, pp. 390-393, 1990.