Binary Co-occurrence Matrix in Image Database Indexing

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
{Iivari.Kunttu, Leena.Lepisto, Ari.Visa}@tut.fi
http://sigwww.cs.tut.fi/
[2] ABB Oy, Paper, Printing, Metals & Minerals, Automation,
P. O. Box 94, FIN-00381 Helsinki, Finland
Juhani.Rauhamaa@fi.abb.com
http://www.abb.com/

Abstract. The use of the second order statistical measures has became popular in the image database indexing and retrieval. Unlike the common approach, image histogram, second order statistics like image correlogram and autocorrelogram consider also the spatial organization of the image colors or gray levels. Recently, correlograms and autocorrelograms have been widely used in the image database indexing. In this paper we present binary co-occurrence matrix, a new statistical measure for image indexing. This measure represents the footprint distribution of the co-occurrence matrix. Compared to image correlogram, this approach provides better retrieval accuracy at lower computational cost. We make retrieval experiments using two industrial image databases. These databases contain images collected from paper and metal manufacturing processes. In the experiments, we compare the retrieval performance of our approach to that of correlograms and autocorrelograms.

1 Introduction

In addition to texture and shape, the distribution of image colors (or gray levels) is an essential feature in content-based image retrieval. Image histogram is a first order statistical measure that has been traditionally used in characterization global color distribution of the image. The benefit of the image histogram is its low computational cost. However, histogram describes only the global distribution of the colors ignoring their spatial organization. This drawback has a remarkable effect on the image retrieval accuracy.

A simple improvement to the color-based image retrieval is the use of second order statistics. The second order statistical measures utilize the spatial organization between the pixel pairs occurring in the image. Correlation-based methods have been used in texture analysis since 1950's. Kaizer [6] was the first who used autocorrelation function to measure texture coarseness. Co-occurrence matrix introduced by Haralick [4] is a correlation-based tool for texture analysis. Correlation function has

J. Bigun and T. Gustavsson (Eds.): SCIA 2003, LNCS 2749, pp. 1090–1097, 2003.

been used also in the field of image retrieval. Huang et al. [5] introduced color correlogram, a measure that describes the spatial correlation of image colors as a function of their spatial distance *d*. In fact, the principle of correlogram is the equal to co-occurrence matrix. The difference between these measures is that whereas co-occurrence matrix uses a single distance *d*, correlogram is calculated for a set of different distances. Because of its computational lightness, Huang et al. preferred autocorrelogram to correlogram in image indexing. Autocorrelogram is a subset of correlogram. It defines the probability of finding identical colors at distance *d*. In [5] retrieval experiments showed that autocorrelogram gives significantly better retrieval results than image histogram. For computational reasons, also Ojala et al. [9] chose autocorrelogram instead of correlogram in image indexing. However, the information carried by correlogram covers the image color content significantly better than autocorrelogram. In [8], we showed retrieval results achieved using correlogram were remarkably better than in the case of autocorrelogram.

In this paper we apply binary co-occurrence matrix in image database indexing. The method was introduced in [7] as a tool for image retrieval without segmentation. The binary co-occurrence matrix is a simple and effective second order statistic for image database indexing. We compare the binary matrix to correlogram-based image retrieval tools. In section two, we present the principles of correlation-based statistical tools and the binary co-occurrence matrix. The retrieval ability of these methods is measured in section three.

The number of digital imaging and image databases in industry has strongly increased during recent years. For example, in process industry, digital imaging solutions are used to control the process and quality. In many cases, these solutions store the image data in image databases. These industrial image databases containing real image data are a challenging retrieval task. In this paper, we use two industrial image databases for testing purposes. The first of these databases is collected from the paper manufacturing process, and it contains 1308 paper defect images. The second testing database is from metal industry. In this database there are 1955 images of defects occurring in the metal surface.

2 Image Database Indexing Using Second Order Statistics

Statistical methods for the analysis of image gray levels or colors are commonly used tools for characterization of the image content. First order statistical methods, like histogram, consider image pixels separately ignoring their spatial relationships. Second- and higher order measures estimate the relationships between two or more pixel values occurring at specific locations relative to each other. In this section, we consider second order statistical measures for image retrieval. In addition to the commonly used methods, we present our approach, binary co-occurrence matrix.

2.1 Statistical Tools for Image Retrieval

Second order statistical measures have traditionally been used in texture analysis. In addition, correlograms and autocorrelograms have also been used in image retrieval. In this part, we present commonly used second order statistical measures.

Image correlogram represents the correlations between the image pixel values. The definition of image correlogram is the following [5] [9]. Let I be an XxY image which comprises of pixels $p(x,y)$. Each pixel has a certain color- or gray level (henceforth level). Let $[G]$ be a set of G levels $g_1 \ldots g_G$ that can occur in the image. For a pixel p, let $I(p)$ denote its level g, and let I_g correspond to a pixel p, for which $I(p)=g$. Let $[D]$ denote a set of fixed distances $d_1 \ldots d_D$. Hence, the number of the distances in this set is D. The correlogram of the image I is defined for level pair (g_i, g_j) at a distance d:

$$\gamma_{g_i,g_j}^{(d)}(I) \equiv \Pr_{p_1 \in I_{g_i}, p_2 \in I} \left[p_2 \in I_{g_j} \middle| |p_1 - p_2| = d \right] \tag{1}$$

which gives the probability that given any pixel p_1 of level g_i, a pixel p_2 at a distance d from the given pixel p_1 is of level g_i. In other words, the correlogram is a matrix that gives the probability of certain level to occur at the distance d from each other. Correlogram is defined for several values of d defined in the set $[D]$. The size of the correlogram-based feature vector is G^2D.

Autocorrelogram [5], [9] is the subset of the correlogram. It captures only the spatial correlation of the identical levels. The autocorrelogram can be defined as:

$$\alpha_g^{(d)}(I) = \gamma_{g,g}^{(d)}(I) \tag{2}$$

and it gives the probability that a pixel p_2, d away from the given pixel p_1, is of level g. In case of the autocorrelogram, the size of the feature vector is GD.

Co-occurrence matrix introduced by Haralick et al. [4] is the basis of the statistical texture analysis. It is a matrix that express the probability of two pixels to occur at certain distance from each other. In fact, co-occurrence matrix is the same as image correlogram defined for a single distance d.

2.2 Binary Co-Occurrence Matrix

A new second order statistical feature to be used in the image retrieval is binary co-occurrence matrix [7]. It is formed by means of the co-occurrence matrix (or a correlogram calculated for a single distance d). In the binary form of the matrix, all the occurrences between the image pixel levels are considered equally. This is done by quantizing the matrix into two levels, “zero” and “non-zero” values. In this way, a binary matrix containing only zeros and ones is formed. The size of the binary co-occurrence matrix is hence G^2.

Table 1. The computational cost based on the length of the feature vectors

FEATURE	FEATURE VECTOR LENGTH
BINARY CO-OCCURENCE MATRIX	G^2
32 gray levels	1024
16 gray levels	256
CORRELOGRAM	G^2D
32 gray levels	4096
16 gray levels	1024
AUTOCORRELOGRAM	GD
32 gray levels	128
16 gray levels	64

Binary co-occurrence matrix has two benefits that make it effective in the image retrieval. First, it is computationally light method compared to the image correlogram (that in our experiments in [8] proved to be the more powerful statistic in image retrieval than image autocorrelogram and histogram). However, the retrieval results of binary co-occurrence matrix are at the same level or better than correlogram. The second reason for the use of binary co-occurrence matrix is the fact that it considers all the correlations occurring in the image equally, which means that in many cases image segmentation can be avoided [7].

2.3 Statistical Measures in Image Database Indexing

Computational cost is an essential property of the indexing methods used in image retrieval. The computational cost of each method is proportional to the length of the feature vector, and therefore short vectors are preferred. The lengths of the feature vectors used in this paper are presented in table 1. In this table, the number of distances (D) is selected to be 4, as in [5] and [9].

Computational cost has been a reason in [5] and [9] for the use of the image autocorrelograms instead of correlograms in the image database indexing. However, in the description of image content, tools based on the whole probability distribution of the image (like correlogram and co-occurrence matrix) are clearly better. In [8] we solved the problem of the computational cost by dividing the database images in the areas of similar color (or gray level). This method is near the principle of color sets presented in [11]. In our approach, this division was made by re-quantizing the color space of the images. This way the number of the image levels G was decreased. Also the quantization of the image generalizes the image content and yield to better retrieval results [8]. Therefore, this quantization is used also in the experiments presented in this paper.

In image retrieval the similarity between the query image **Q** and the database image **I** is measured by distance metrics. In [5] and [9], the distance measure between the autocorrelograms is L_1 norm [2]:

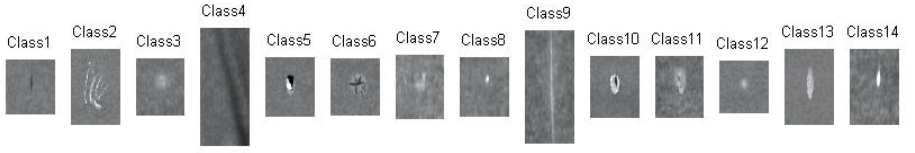

Fig. 1. An example of each paper defect image class in testing database I

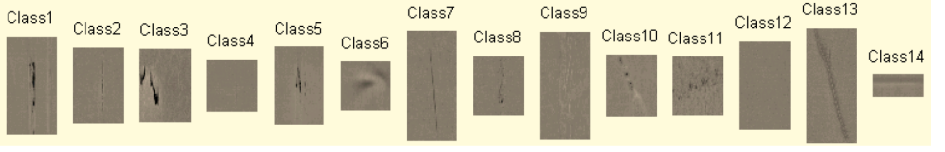

Fig. 2. An example of each metal defect image class in testing database II

$$L_1 = \sum_{i=1}^{N} \left| \alpha_{g\mathbf{Q}}^{(d)}(i) - \alpha_{g\mathbf{I}}^{(d)}(i) \right| \tag{3}$$

We use the same distance measure also in case of the correlograms. In case of binary co-occurrence matrices, binary distances are required. When comparing two binary matrices, $\mathbf{B_1}$ and $\mathbf{B_2}$, let $n_{1,1}$ denote the number of the elements, whose value is 1 in both matrices. In a similar way, $n_{1,0}$, $n_{0,1}$ and $n_{0,0}$ denote numbers of matrix elements, which have values 1 and 0, 0 and 1, 0 and 0, respectively. Jaccard coefficient [3] is a popular similarity measure for binary data. This coefficient is defined as:

$$S_J = \frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}} \tag{4}$$

3 Experiments

In this section we test our approach to the correlogram-based image retrieval using real industrial image databases. For testing purposes we had two sets of defect images. In testing database I, the defects were collected from the paper web using a paper inspection system [10]. The objects in the images were typical paper surface defects. The test set consisted of 1308 paper defects, which represented 14 defect classes so that each class consisted of 32-100 defect images. An example image of each paper defect image class is presented in figure 1. The second test set, testing database II, there were 1955 metal surface defect images. Also in this case, there were 14 defect classes (figure 2). Each class contained 100-150 metal defect images. In both databases, the images were intensity images containing 256 gray levels. The image size had also strong variations (dimension of the image varied from 100 to 2000 pixels).

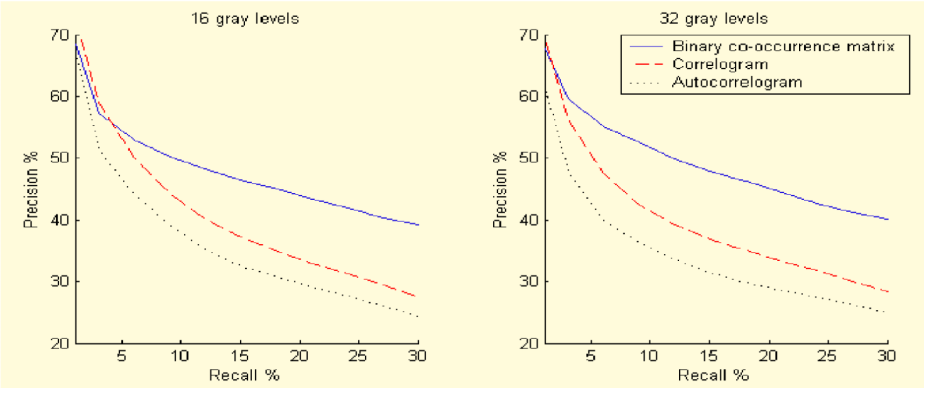

Fig. 3. Average retrieval performance of the features in case of paper defect images in testing database I

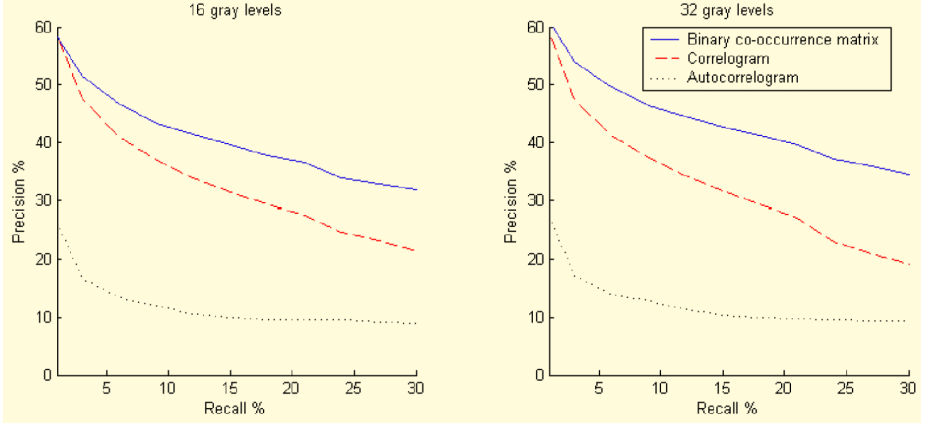

Fig. 4. Average retrieval performance of the features in case of metal defect images in testing database II

We calculated the correlograms, autocorrelograms as well as the binary co-occurrence matrices for the database images quantized to 32 and 16 gray levels. In the calculation of the autocorrelogram and the correlogram we used the set of distances $[D]=\{1,3,5,7\}$, which is the same as in [5] and [9]. Both databases were indexed using these features. The purpose of the retrieval experiments was to test the retrieval ability of each feature. The retrieval experiments were made using *leaving one out* method [3]. In this method, each image in turn is left out from the test set and used as a query image, whereas the other images in the test set form a testing database. In the queries, the nearest images to the query image are retrieved based on their feature vectors. The performance of the retrieval was measured by calculating a *precision*

versus recall curve [1] for each query. If $|A|$ is the number of all retrieved images, $|R|$ is the number of query class images in the whole testing database and $|Ra|$ is the number of retrieved query class images, precision and recall can be defined in the following way [1]:

$$\text{Precision} = \frac{|Ra|}{|A|} \tag{5}$$

$$\text{Recall} = \frac{|Ra|}{|R|} \tag{6}$$

The retrieval performance of each feature can be presented by calculating the average precision-recall curve for each query.

In the retrieval experiments, the similarity measure for the binary co-occurrence matrices was the Jaccard coefficient. For the correlograms and the autocorrelograms, L_1 norm was used. Figures 3 and 4 present the average precision-recall curves for the images in both testing databases.

4 Discussion

In this paper we presented a new approach to the statistical image retrieval. Our method, binary co-occurrence matrix is a simple tool for the characterization of the gray level distributions of the database images. The experimental results presented in figures 3 and 4 show that the binary co-occurrence matrix is an effective method in image retrieval. Compared to the correlogram and autocorrelogram, our method gives clearly better results in retrieval accuracy. Binary co-occurrence matrix is also computationally lighter method than image correlogram.

For testing purposes, we had two industrial image databases. The testing databases contained 1308 paper defect images and 1955 images of metal defects. These kinds of industrial databases provide a good opportunity to test the retrieval methods with real image data. On the other hand, retrieval of the defect images is a quite demanding task. This is because some defect classes are very similar to each other and also overlapping. In these classes, the retrieval results could be improved by using some shape descriptors together with the statistical tools.

In this work, the testing databases contained only gray level images. Our method, binary co-occurrence matrix could be applied also to the classification and retrieval of color images. In that case the color quantisation can be made using tools presented in [11]. This could be a subject of further studies in the field of statistical image retrieval methods.

Acknowledgment

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

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, Addison Wesley, New York (1999)
2. Duda, R.O., Hart, P.E., Stork, .G.: Pattern Classification. 2nd ed. John Wiley (2001)
3. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press, Massachusetts, USA (2001)
4. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-3, 6 (1973)
5. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.-J., Zabih, R.: Image indexing Using Color Correlograms. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Juan, Puerto Rico (1997) 762-768
6. Kaizer, H.: A Quantification of Textures on Aerial Photographs. Tech. Note No. 121, A.69484. Boston University Research Laboratories, Boston University (1955)
7. Kunttu, I., Lepistö, L., Rauhamaa, J., Visa, A.: Image Retrieval without Segmentation. Proceedings of 10th Finnish AI Conference. Oulu, Finland (2002) 164-169
8. Kunttu, I., Lepistö, L., Rauhamaa, J., Visa, A.: Image Correlogram in Image Database Indexing and Retrieval. Proceedings of 4th European Workshop on Image Analysis for Multimedia Active Services, London, UK, (2003) 88-91
9. Ojala, T., Rautiainen, M., Matinmikko, E., Aittola, M.: Semantic Image Retrieval with HSV Correlograms. Proceedings of 12th Scandinavian Conference on Image Analysis. Bergen, Norway (2001) 621-637
10. Rauhamaa, J., Reinius, R.: Paper Web Imaging with Advanced Defect Classification. TAPPI Technology Summit, Atlanta, Georgia (2002)
11. Smith, J.R., Chang, S.F.: Tools and Techniques for Color Image Retrieval. Storage and Retrieval for Image and Video Databases IV, SPIE Proceedings, Vol. 2670 (1996) 1630-1639