Performance Analysis in Content-based Retrieval with Textures*

Kun Xu[(1)], Bogdan Georgescu[(2)], Dorin Comaniciu[(3)], Peter Meer[(1)]
(1)Electrical and Computer Engineering Department
(2)Department of Computer Science
Rutgers University, Piscataway, NJ, 08855-0909, USA
(3)Imaging & Visualization Department, Siemens Corproate Research
755 College Road East, Princeton, NJ 08540, USA
kunx, georgesc, comaniciu, meer@caip.rutgers.edu

Abstract

The features employed in content-based retrieval are most often simple low-level representations, while a human observer judges similarity between images based on high-level semantic properties. Using textures as an example, we show that a more accurate description of the underlying distribution of low-level features does not improve the retrieval performance. We also introduce the simplified multiresolution symmetric autoregressive model for textures, and the Bhattacharyya distance based similarity measure. Experiments are performed with four texture representations and four similarity measures over the Brodatz and VisTex databases.

Keywords: content-based retrieval, texture description, similarity measure.

1. Introduction

Retrieval from a database of images (video sequences) by finding semantic similarities with the visual information contained in the query, is a task of great practical interest today. Numerous systems were built and some are even enjoying commercial success. See [1] for a comprehensive review.

The similarity measure between a query image and the images in the database is usually computed employing low-level features associated with salient regions: color, texture, shape etc. These features provide only a crude representation of the image and most of the semantic information, the very content which distinguishes an image from other types of information, is lost.

In this paper we investigate performance bounds in content-based image retrieval due to the inadequacy of the employed feature representation. We chose *texture* as a case study since two standard (partially overlapping) databases are available [12, 13], and the problem of texture modeling while relative simple and well defined, still exhibits the pitfalls of inadequate representations.

2. Images and Texture Homogeneity

The Brodatz [12] and VisTex [13] databases are frequently employed in texture studies. The latter contains a subset of the former. Brodatz has 112 while VisTex contains 132, 512x512 gray level images.

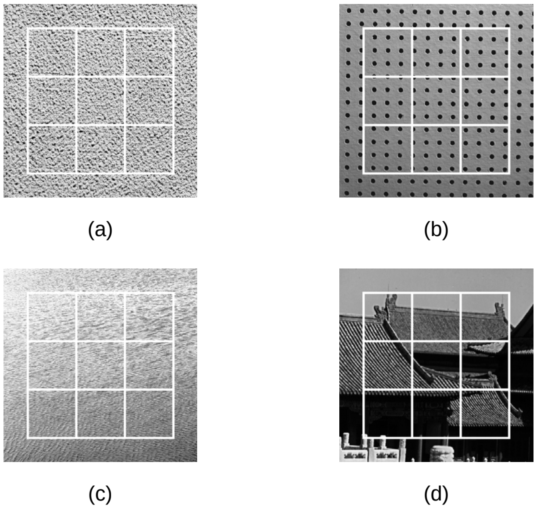

(a) (b) (c) (d)

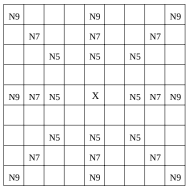
Figure 1. Examples of classes. (a) D57 Handmade Paper (Brodatz). (b) Tile.0007 (VisTex). (c) D38 Water (Brodatz). (d) Tile.0005 (VisTex).

A class is defined by dividing the 384x384 central part of each image into nine nonoverlapping 128x128 images [5]. Thus the Brodatz database contains 1008 and the VisTex database 1188 images.

The three main perceptual properties of textures: periodicity, directionality and randomness [5] are easy to recognize by humans but elusive when to be described quantitatively by a machine. For example, the resolution of the analysis is a crucial parameter since the periodicity of a texture is directly related to the size of the texture element (texel). While it is possible to have an approximate estimate of the texel size [4], in Section 3 only texture representation methods which use the same window sizes for the entire database: the multiresolution simultaneous autoregressive (MRSAR) model [6], and the Gabor filter bank [3], are considered.

Let examine the nine 128x128 images associated with the same class. Since the original 512x512 images included into the texture databases *were classified as texture by a human observer*, it is not unexpected to find significant differences between separate regions of the same image. In Figure 1 examples of decreasing class homogeneity are shown. The nine images in Figure 1a are very similar, while it will be difficult to retrieve using as query the image in the upper left cor-

*The research was supported by the NSF under the grants IRI 95-30546 and IRI 96-18854.



N9				N9				N9
	N7			N7			N7	
		N5		N5		N5		
N9	N7	N5		X		N5	N7	N9
		N5		N5		N5		
	N7			N7			N7	
N9				N9				N9

Figure 2. Neighborhood definitions for the three resolutions of the MRSAR model.

ner of Figure 1d the other eight images belonging to the same class.

3. Feature Extraction Methods

In the spatial domain textures are characterized by a two step procedure. First, a window is slided across the image and at each location the local structure is represented by a vector. Next, the mean and the covariance of these vectors is used as the texture representation of the entire image. To reduce the artifacts due to the difference between the size of the window and the texel, the procedure is repeated for several window dimensions. The final representation is the concatenation of the outputs of individual procedures.

3.1. Gabor Filters

In the spatial domain the 2D Gabor functions are complex sinusoidal gratings modulated by 2D Gaussian functions. In the spatial frequency domain they correspond to 2D bandpass filters. A typical Gabor filter bank design is described in [3]. The frequency domain filtering is equivalent to applying 24 spatial filters of increasing sizes, and thus a 24-dimensional feature vector is associated with every pixel in the image.

3.2. MRSAR Method

The MRSAR method models the texture as a second-order noncausal Markov random field. In [6] the four parameters of the underlying autoregressive model are estimated independently at three resolutions using windows of size 5x5, 7x7 and 9x9 [2]. For each resolution k the model is defined as

$$g(i,j) = \sum_{(m,n)\in\mathcal{N}_k} a_k(m,n)g(i-m,j-n)+n_k(i,j) \quad (1)$$

where $\mathcal{N}_k$ is the employed neighborhood of pixel (i,j) at resolution k, see Figure 2, $g(\cdot,\cdot)$ the gray level values in the image and $n_k(i,j)$ the error term associated with the model. A symmetric model is assumed with $a_k(n,m) = a_k(-n,-m)$ for all k. Together with the standard deviation σ_k of the error term, at each resolution five parameters are estimated and after concatenation a 15-dimensional feature vector is obtained.

The least squares estimations are carried out in a large 21x21 window slided across the image with two pixel steps. Careful analysis of the MRSAR procedure reveals that

- At each resolution the estimation process in the 21x21 window integrates together representations of widely different local structures.
- The noise process driving the AR model is independent of the resolution.
- The noise standard deviation σ_k is one to two orders of magnitude larger than the autoregressive coefficients $a_k(n,m)$.

Based on the above observations the original MRSAR procedure can be simplified. The texture in the 21x21 region is represented by a single, 12-dimensional symmetric autoregressive model,

$$g(i,j) = \sum_{(m,n)\in\mathcal{N}} a(m,n)g(i-m,j-n)+n(i,j) \quad (2)$$

where now $\mathcal{N}$ includes all the locations marked in Figure 2. Together with the standard deviation of the noise σ, thus the texture is represented by a 13-dimensional vector. This representation will be called ORSAR (one resolution SAR). As will be shown, the retrieval performance is not significantly affected by using the ORSAR representation.

4. Similarity Measures and Performance Assessment

The ensemble of locally defined texture representations (feature vectors) is traditionally characterized by mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{C}$. Employing only the first two moments of the ensemble, the distribution is implicitly assumed of being unimodal and normal.

Given the query image, described as $(\boldsymbol{\mu}_q, \boldsymbol{C}_q)$, an often used similarity measure is its Mahalanobis distance from the entries in the database $(\boldsymbol{\mu}_d, \boldsymbol{C}_d)$

$$m = (\boldsymbol{\mu}_q - \boldsymbol{\mu}_d)^T \boldsymbol{C}^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_d) \ . \quad (3)$$

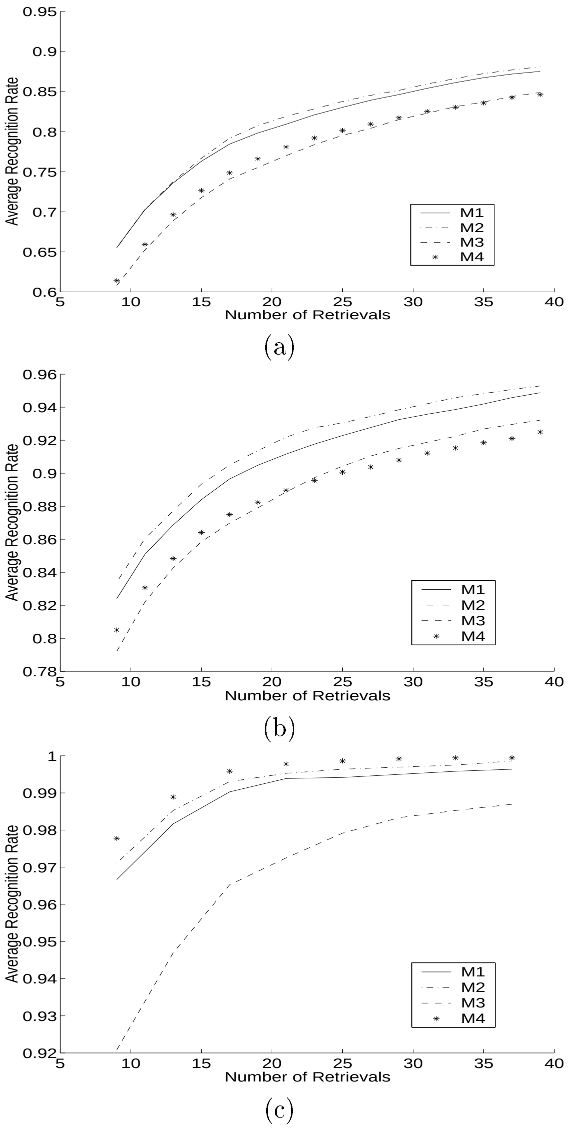
The covariance of the query $\boldsymbol{C} = \boldsymbol{C}_q$ is used to define the underlying metric. However, the Mahalanobis distance fails when the two distributions differ only by their second order statistics. The Bhattacharyya distance [7, p.99]

$$\begin{aligned} b &= \frac{1}{4}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_d)^T(\boldsymbol{C}_q + \boldsymbol{C}_d)^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_d) \\ &\quad + \frac{1}{2}\ln\frac{\left|\frac{\boldsymbol{C}_q+\boldsymbol{C}_d}{2}\right|}{\sqrt{|\boldsymbol{C}_q||\boldsymbol{C}_d|}} \qquad (4) \end{aligned}$$

on the other hand, takes all the available information into account. Recently, we have shown an efficient way to compute the Bhattacharyya distance exploring the special structure of most feature spaces [9].

To compare the performance of the three feature representation (Gabor, MRSAR, ORSAR) and two similarity measures (Mahalanobis, Bhattacharyya), in Figure 3 the average recognition rate is plotted against the number of retrievals for two databases (Brodatz, VisTex). The databases have over 1000 entries, but in practical situations only the quality of the first few (say) 40 retrievals is of interest.

The Bhattacharyya based similarity measure always outperformed the traditional Mahalanobis based measure. The Gabor filter has slightly better performance



(a)

(b)

(c)

Figure 3. The retrieval performance. (a) VisTex database. (b) Brodatz database. (c) hBrodatz database. Employed representation/similarity measure (feature space dim). M1:ORSAR/Bhatt(13). M2:MRSAR/Bhatt(15). M3:MRSAR/Maha(15). M4:Gabor/Bhatt(24).

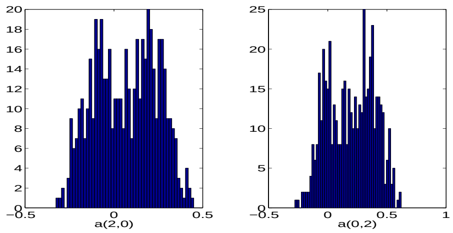


Figure 4. Marginal histograms of two autoregressive coefficients.

than the MRSAR (with Mahalanobis similarity measure) as was also reported in [3]. The VisTex database with less homogeneous classes yielded a lower retrieval rate.

The effect of class inhomogeneity on the retrieval performance can be seen by defining the homogeneous hBrodatz database. For this database 50 images were selected from the original Brodatz set. The example in Figure 1a is typical for hBrodatz. The retrieval performance shown in Figure 3c. In comparison to Figure 3a the recognition rates are shifted upward with at least 0.1, except the method M4 (Gabor/Bhatt) which became the best. This change in performance ranking may explain why often texture segmentation studies cannot find a universally optimal feature representation method [10, 11].

In the sequel will focus on the ORSAR representation with Bhattacharyya similarity measure to show the intrinsic limitations of nonsemantical, exclusively low-level features based retrievals.

5. Semiparametric and Nonparametric Similarity Measures

The feature vector distribution derived from an image is not necessarily unimodal, especially if the texel and the window of analysis have comparable sizes. A 13-dimensional distribution is difficult to visualize, and in Figure 4 two marginal distributions are shown for an image of the class in Figure 1b. Note the strong bimodality. We can ask the question: will a more accurate description of the feature distribution improve the retrieval performance?

To investigate this issue the Bhattacharyya distance between two *arbitrary* distributions $q(\boldsymbol{x})$ and $d(\boldsymbol{x})$ will be employed [7, p.99]

$$B(\boldsymbol{q},\boldsymbol{d}) = -\ln\int\sqrt{q(\boldsymbol{x})d(\boldsymbol{x})}d\boldsymbol{x}\ . \qquad (5)$$

5.1. Semiparametric Representation

The feature vector distribution is represented as the mixture of M multivariate normals. For computational considerations M was kept small, M = 4 in our experiments. Using a simple ISODATA procedure [8, p.98] the feature space is first clustered into M clusters, which are approximated by a mean vector and a covariance matrix. Then

$$\begin{aligned} B(\boldsymbol{q},\boldsymbol{d}) &= -\ln\int\sqrt{\left(\sum_{i=1}^{M}\frac{n_i}{N}q_i(\boldsymbol{x})\right)\left(\sum_{j=1}^{M}\frac{n_j}{N}d_j(\boldsymbol{x})\right)}\ d\boldsymbol{x} \\ &\approx -\ln\sum_{i=1}^{M}\frac{\sqrt{n_i n_{j_i}}}{N}\int\sqrt{q_i(\boldsymbol{x})d_{j_i}(\boldsymbol{x})}d\boldsymbol{x} \\ &\approx -\ln\sum_{i=1}^{M}\frac{\sqrt{n_i n_{j_i}}}{N}e^{-b(\boldsymbol{q}_i,\boldsymbol{d}_{j_i})} \qquad (6) \end{aligned}$$

where n_i is the number of points belonging to the i-th multivariate normal, and $j_i = \arg\min_j b(\boldsymbol{q}_i,\boldsymbol{d}_j)$, with b computed as in (4).

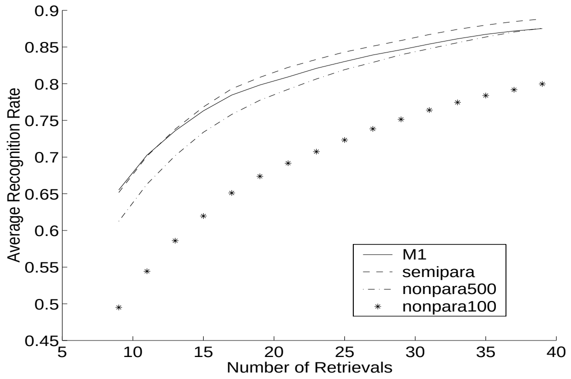


Figure 5. The retrieval performance for the VisTex database using different methods.

In Figures 5 and 6 the retrieval performance using the semiparametric representation are shown. The performance does not change significantly.

5.2. Nonparametric Representation

To avoid the artifacts introduced by the mixture of Gaussians, the distance (5) can be evaluated directly. First however, the feature space has to be reduced to N data points. We used N = 100 and 500 in the experiments. The retained points correspond to the N densest regions, each containing the same number of points in the original set. These regions are delineated by analysing the nearest neighbor distances.

Next, the retained point sets are scaled to be within the unit 13-dimensional hypercube using as scaling factors the largest value in the database along each dimension. The terms $q(\boldsymbol{x}) \cdot d(\boldsymbol{x})$ are computed by centering a $h < 1$ size hypercube on the two points of a pair of nearest neighbors in the *combined* $\boldsymbol{q}, \boldsymbol{d}$ point sets, and finding the volume of the intersection. The optimal h was determined by Bayesian inference from the histograms of nearest neighbor distances for nine representative classes from the database. For example, for $N = 500$ and the Brodatz database, $h = 0.22$. The nonparametric similarity measure is then defined as

$$B(\boldsymbol{q}, \boldsymbol{d}) = 1 - \frac{1}{N} \sum_{i \in \boldsymbol{q}, j_i \in \boldsymbol{d}} V_{i,j_i} \qquad (7)$$

where V_{i,j_i} is the volume of overlap between the h-sized hypercubes centered on the points i in point set $\boldsymbol{q}$ and j_i its nearest neighbor in $\boldsymbol{d}$. Note that V_{i,j_i} can be zero.

The retrieval performance (Figures 5 and 6) improves with N, however never reaches that of the parametric ORSAR approach. The sensitivity of nonparametric representations in such high dimensional spaces is probably more detrimental than the simple, unimodal description by ORSAR.

6. Conclusion

We conclude based on the experiments presented that the main factor in limiting texture retrieval performance is not the inaccurate description of the feature distribution, but the nonhomogeneity of the images within a class. Extrapolating this observation to the general content-based retrieval problem, our findings suggest the importance of including semantic descriptions, however simplistic they are. These descriptions should attempt to capture more global invariant characteristics than those represented by the low-level features.

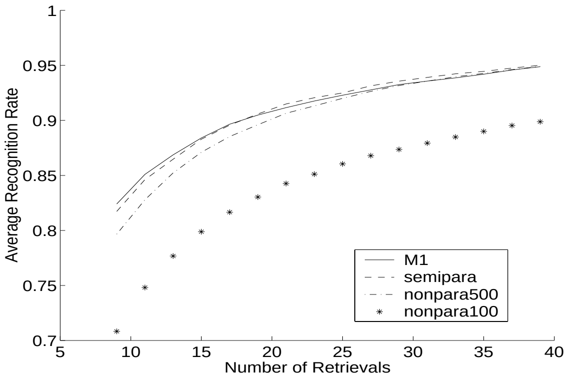


Figure 6. The retrieval performance for the Brodatz database using different methods.

References

[1] S. Antani, R. Kasturi, R. Jain, "Pattern recognition methods in image and video databases: past, present and future", *Advances in Pattern Recognition, Lecture Notes in Comp. Science*, Springer, Vol. 1451, 31–53, 1998.

[2] R.W. Picard, T. Kabir, F. Liu, "Real-time recognition with the entire Brodatz texture database", *IEEE Conf. on CVPR*, New York, 1993, 638–639.

[3] B.S. Manjunath, W.Y. Ma, "Texture features for browsing and retrieval of image data", *IEEE Trans. on PAMI*, Vol. 18, 837–842, 1996.

[4] A. Khotanzad, J.-Y. Chen, "Unsupervised segmentation of textured images by edge detection in multidimensional features", *IEEE Trans. on PAMI*, Vol. 11, 414–420, 1989.

[5] F. Liu, R.W. Picard, "Periodicity, directionality, and randomness: Wold features for image modeling and retrieval", *IEEE Trans. on PAMI*, Vol. 18, 722–733, 1996.

[6] J. Mao, A.K. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models", *Pattern Recognition*, Vol. 25, 173–188, 1992.

[7] K. Fukunaga, "*Introduction to Statistical Pattern Recognition*", Second Edition, Academic Press, 1990.

[8] A.K. Jain, R.C. Dubes, "*Algorithms for Clustering Data*", Prentice Hall, 1988.

[9] D. Comaniciu, P. Meer, K. Xu, D. Tyler, "Retrieval performance improvement through low rank corrections", *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, Fort Collins CO, June 1999, 50–54.

[10] P.P. Ohanian, R. C. Dubes, "Performance evaluation for four classes of textural features", *Pattern Recognition*, Vol. 25, 819–833, 1992.

[11] T. Randen, J.H. Husøy, "Filtering for texture classification: A comparative study", *IEEE. Trans. on PAMI*, Vol. 21, 291–310, 1999.

[12] P. Brodatz, *"Textures: A Photographic Album for Artists and Designers"*, Dover, New York, 1966.

[13] Vision Texture Database, MIT Media Lab,
`www-white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html`