# Non–parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval *

Jan Puzicha Thomas Hofmann Joachim M. Buhmann
Institut für Informatik III, University of Bonn
Römerstraße 164, D–53117 Bonn, Germany
http://www-dbv.cs.uni-bonn.de

## Abstract

*In this paper we propose and examine non–parametric statistical tests to define similarity and homogeneity measures for textures. The statistical tests are applied to the coefficients of images filtered by a multi–scale Gabor filter bank. We will demonstrate that these similarity measures are useful for both, texture based image retrieval and for unsupervised texture segmentation, and hence offer an unified approach to these closely related tasks. We present results on Brodatz–like micro–textures and a collection of real–word images.*

## 1 Introduction

Color, Shape, Motion and Texture are the basic modes to describe low–level image content and have been used to both, measure similarity of images, and segment images into homogeneous regions. The definition of suitable similarity and homogeneity measures for these modes is a fundamental task in many important applications, ranging from vision–guided autonomous robotics and remote sensing to medical diagnosis and similarity–based retrieval in large image databases such as the QBIC system [1] or the MIT Photo–book [8].

With the restriction to a set of known textures, retrieval and segmentation problems are essentially reduced to a supervised classification task, which is amenable for standard techniques from pattern recognition and statistics. As opposed to supervised methods which rely on labeled data to learn decision boundaries in some appropriate feature space, the central topic of unsupervised segmentation is concerned with the weaker notion of *texture proximity*, based on a general (not class– or texture–specific) *similarity measure*. Inspired by the supervised approach, the majority of unsupervised methods formulate the retrieval and segmentation problems in a feature–based fashion. This conception inevitably leads to the difficult problem of specifying a metric in the utilized feature space which appropriately represents visual dissimilarities between textures [5, 6, 3]. In contrast to this widely appreciated approach, we follow the ideas of [2, 7] and advocate *non–parametric statistical tests* to measure texture similarity. Statistical tests have the advantage to be applicable without parametric assumptions about the underlying pixel distribution. This guarantees the similarities to be assessable in terms of statistical significance, but avoids statistical parameter estimation.

There exists a tight relationship between similarity–based image retrieval and unsupervised texture segmentation. Image retrieval often requires to select those (parts of) images in a database which are most similar to a given query image, while the goal of image segmentation is to partition a given image into maximally homogeneous regions. Therefore these tasks are closely related to similarity measures, since homogeneity can be defined as the average similarity between pairs of local texture patches within a region. While similarity–based retrieval is straightforward for a given measure, we model the texture segmentation problem as a combinatorial optimization problem specified by a *pairwise data clustering* objective function. The choice of a suitable objective function is crucial, and has been motivated by certain invariance properties [4], i.e., linear transformations of the dissimilarities. Experimental evidence, too, suggests that these invariant functions are superior to the standard graph partitioning cost function utilized in [2].

## 2 Image Representation

The differential structure of an image $I(\vec{x})$ is completely extracted by the classical scale–space representation. But in many applications it is convenient to use filters which are tuned to the features of interest, e.g., a particular spatial frequency $\vec{k}$. The tuning operation can be formalized and, in

*Supported by the German Research Foundation (DFG # BU 914/3–1) and by the Federal Ministry for Education, Science and Technology (BMBF # 01 M 3021 A/4). It is a pleasure to thank the MIT Media Lab for providing the VisTex database.

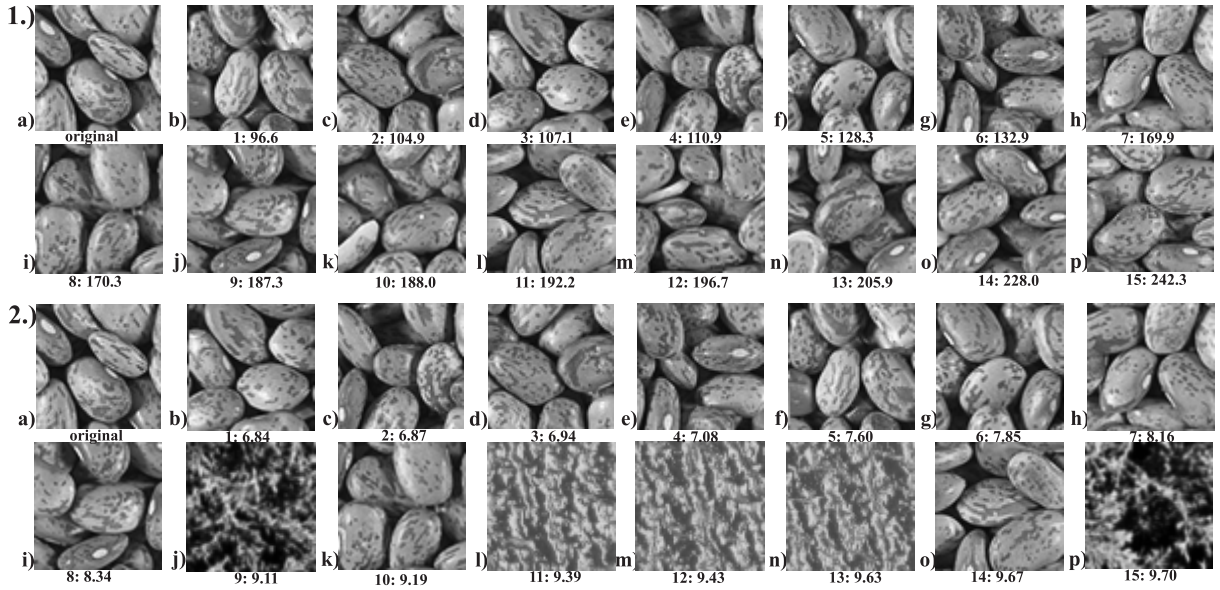


**Figure 1.** Example retrieval **1.)** for the $\chi^2$–statistic and **2.)** for WMV. Images are 64x64 pixels each. The database consists of 16 samples of each of 40 reference textures. The image captions depict the retrieval rank $r$ and the measured distance $D$.

the case of frequency tuning, yields the family of complex Gabor filters

$$G(\vec{x},\sigma,\vec{k}) \quad = \quad \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\vec{x}^t\vec{x}/2\sigma\right) \exp\left(i\vec{k}^t\vec{x}\right) \quad (1)$$

Gabor filters perform a local Fourier analysis and exhibit excellent discrimination properties over a broad range of textures [5]. The Gabor multi-scale image representation is especially useful for unsupervised texture processing, where little is known a priori about the characteristic frequencies of occurring textures. In this work, an image $I$ is represented by a set of filtered images $I_r$, defined by the modulus of the filter outputs, $I_r(\vec{x}) = |I(\vec{x}) * G(\vec{x},\sigma_r,\vec{k}_r)|$. We have chosen $12$ Gabor filters at $4$ orientations and $3$ scales separated by octaves with $\sigma_r = 3/||k_r||$ in our experiments. Although we are convinced of the advantages offered by the Gabor representation for defining meaningful and robust similarity measures, the techniques presented in the sequel can be easily adapted to other feature sets.

## 3 Similarity Measures and Image Retrieval

To evaluate the dissimilarity between two textured images or image regions $I$ and $J$, a statistical test $D$ is applied to the distribution of Gabor coefficients, either independently for each pair $I_r$ and $J_r$ of filtered images, or to the joint distribution of coefficients in all channels. If the significance that both samples were drawn from the same distribution is high (low), their dissimilarity $D(I,J)$ is judged low (high). More formally, we denote by $F_r(\cdot;I)$ the empirical *probability distribution function* (PDF) of Gabor coefficients in the filtered image $I_r$ of size $L \cdot M$,

$$F_r(t;I) = |\{\vec{x} \in I_r : \; I_r(\vec{x}) \leq t\}| \,/\, (L \cdot M) \qquad (2)$$

and by $f_r^k(I) = f_r(t_k;I)$ the empirical density (histogram) obtained by suitable binning $t_0 < t_1 < \cdots < t_K$. The generalization to the multidimensional case for joint distributions is straightforward and is omitted for brevity.

Several non–parametric test statistics are empirically investigated:

- The *Kolmogorov–Smirnov distance* as originally proposed in [2]. It is defined as the maximal distance of the PDFs, $D^r(I,J) = \max_t |F_r(t;I) - F_r(t;J)|$.
- A *statistic of the Cramer/von Mises type* defined as the Euclidean distance of the PDFs, $D^r(I,J) = \int (F_r(t;I) - F_r(t;J))^2 dt$, which is rescaled by the coefficient variance to achieve comparable statistics for all channels.
- The $\chi^2$*–statistic* $D^r(I,J) = \sum_{k=1}^{K} \frac{\left(f_r^k(I) - f_r^k\right)^2}{\hat{f}_r^k}$, where $\hat{f}_r^k = [f_r^k(I) + f_r^k(J)]/2$.
- The empirical Jeffrey–divergence defined by $D^r(I,J) = \sum_k f_r^k(I) \log \frac{f_r^k(I)}{\hat{f}_r^k} + f_r^k(J) \log \frac{f_r^k(J)}{\hat{f}_r^k}$, which in contrast to the Kullback–Leibler divergence suggested in [7] is numerically stable, symmetric and robust with respect to noise and the size of histogram bins.
- The Weighted–Mean–Variance (WMV) proposed in [6]. For empirical means $\mu_r(I), \mu_r(J)$ and standard

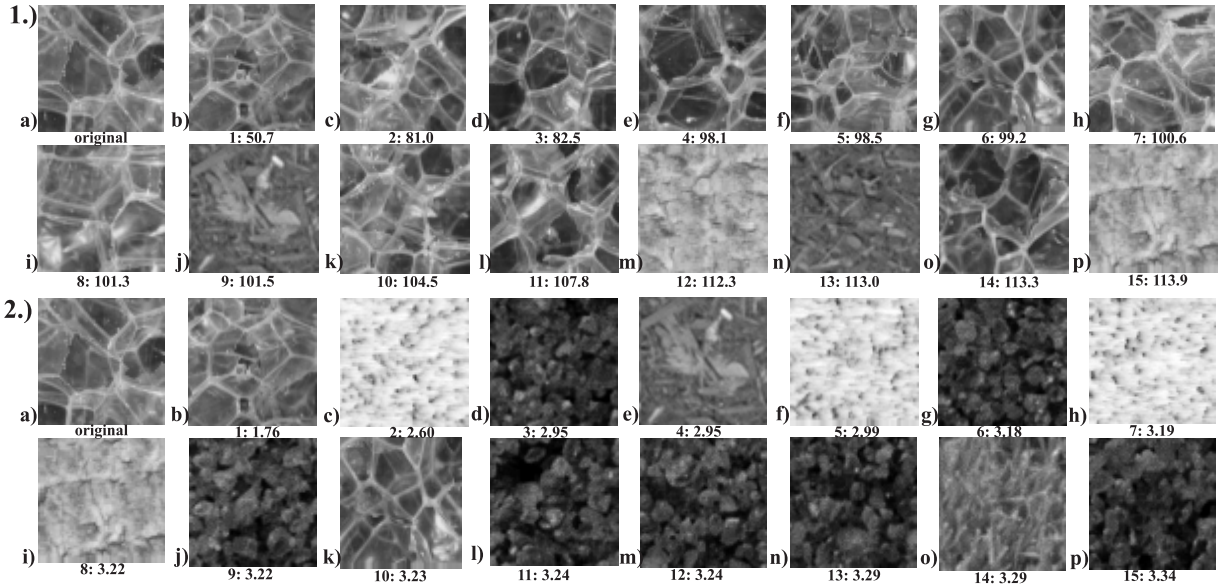


**Figure 2.** Example retrieval for the same database as in Fig. 1. The WMV–measure (2) performs significantly worse and does not correspond well to visual similarity.

deviations $\sigma_r(I), \sigma_r(J)$ the distance is defined by

$$D^r(I,J) = \frac{|\mu_r(I) - \mu_r(J)|}{|\sigma(\mu_r)|} + \frac{|\sigma_r(I) - \sigma_r(J)|}{|\sigma(\sigma_r)|}, \quad (3)$$

where $\sigma(\cdot)$ denotes an estimate of the standard deviation of the respective entity. This measure based on a Gabor filter image representation has outperformed several other parametric models [6].

While a statistical test is a reliable measure to judge the dissimilarity of two sample sets in a single channel, the question arises how to combine the independently evaluated comparisons. We have investigated Minkowski norms $D(I,J) = \sum_r (D^r(I,J))^p$, including the limiting case of the maximum norm ($p = \infty$) utilized in [2]. The Minkowski norm is less sensitive to differences in single channels for small $p$, while large $p$ avoid the 'curse of dimensionality'. For a medium number of $10 - 30$ dimensions, the choice of $p = 1$ empirically showed the best performance.

Once a dissimilarity measure $D(I,J)$ has been specified, the retrieval for a query $J$ is obtained by sorting all database images $I^{(n)},\ 1 \le n \le N$ in ascending order of the dissimilarities $D(I^{(n)}, J)$. Either a fixed number of matches or all matches with dissimilarity below a predefined threshold are displayed.

## 4 Unsupervised Texture Segmentation

In our approach, textured image segmentation is formulated as a combinatorial optimization problem belonging to the class of *partitioning* or *clustering problems*, where a set of $N$ sites $\vec{x}_i$ is mapped to a set of texture labels. For notational convenience we introduce an indicator function representation $M_{i\nu} \in \{0,1\}$ denoting that site $i$ is mapped to label $\nu$. If the number of distinctive textures $K$ is known a priori, a segmentation is summarized in terms of a Boolean assignment matrix $\mathbf{M} \in \mathcal{M}$, with

$$\mathcal{M} = \left\{ \mathbf{M} \in \{0,1\}^{N \times K} \ : \ \sum_{\nu=1}^{K} M_{i\nu} = 1,\ 1 \le i \le N \right\}.$$

For the image segmentation problem, we evaluate proximities $D_{ij}$ between pairs of image windows located at positions $\vec{x}_i$ and $\vec{x}_j$. For simplicity we consider squared areas around the center point, where the size of the window is chosen proportional to the scale parameter $\sigma_r$ of the Gabor filter [5]. The data clustering cost function thus has to rely on the proximity matrix $D = (D_{ij}) \in \mathbb{R}^{N \times N}$. While vector–valued data with a fixed number of features scales linear with the number of sites $N$, pairwise comparison results in a scaling with $N^2$. Yet, it is obvious, that a complete proximity matrix possesses a significant inherent redundancy. To guarantee computational efficiency, the calculation of dissimilarities is restricted to positions $\vec{x}_i$ on a regular sub–lattice of the image. Moreover, comparisons are only made with a substantially reduced set of pairs $(\vec{x}_i, \vec{x}_j)$. This subset is specified in terms of a *neighborhood system* $\mathcal{N} = (\mathcal{N}_i)_{i=1,\ldots,N}$, $\mathcal{N}_i \subset \{1,\ldots,N\}$, which is an irreflexive and symmetric binary relation. Following [2], we define the *neighborhood* $\mathcal{N}_i$, $|\mathcal{N}_i| \ll N$ of a site $\vec{x}_i$ to consist of the four connected neighborhood in the image and a larger number of random neighbors. The main problem from a modeling perspective is the specification of an objective function $\mathcal{H} : \mathcal{M} \to \mathbb{R}$ to assess the qual-

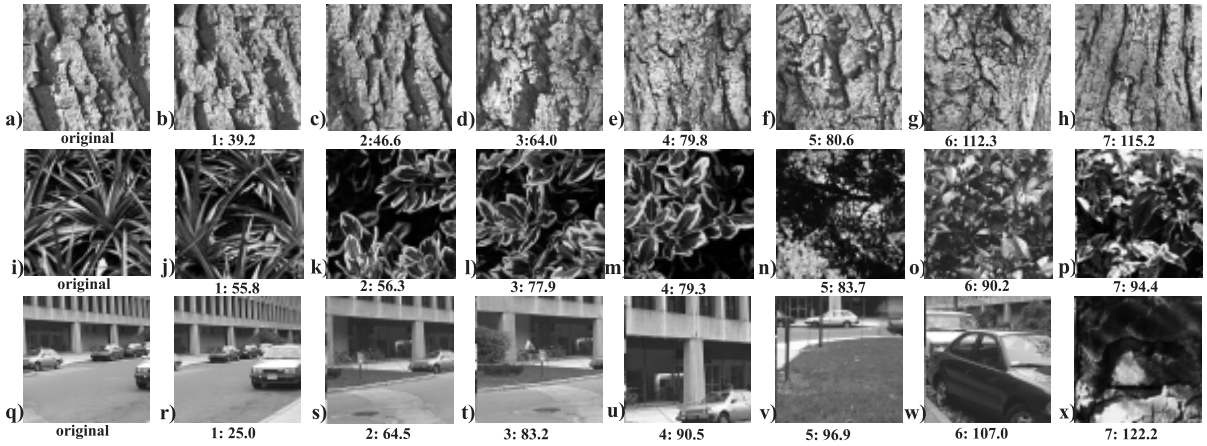


**Figure 3.** Example image retrievals: (a)–(h) bark, (i)–(p) plant, (q)–(x) street scene. The database consists of 668 images with 64x64 pixels each ranging from homogeneous textures to inhomogeneous textured outdoor scenes.

ity of image partitionings $\mathbf{M}$. In this work, we focus on functions measuring intra–cluster compactness, which only depend on the homogeneity of a segment. Additionally, we demand invariance of $\mathcal{H}$ with respect to linear transformations of the dissimilarity matrix. This has the advantage not to introduce a dependency on the minimum of the dissimilarity function. Indeed, it has been noticed [2] that the data have to be shifted appropriately in order to keep the right balance between negative and positive contributions, when using the standard graph–partitioning function $\mathcal{H}^{\mathrm{gp}}(\mathbf{M}) = \sum_{\nu=1}^{K} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} M_{i\nu} M_{j\nu} D_{ij}$. However, if an image contains many different textures, it is often impossible to globally shift the data, such that all textures are well–discriminated. Furthermore it is impossible to select a correct shift for a larger set of images [4].

Taking the invariance properties as our major guideline for an axiomatic derivation of clustering functions [4], we arrive at the following objective function

$$\mathcal{H}(\mathbf{M}) = \sum_{\nu=1}^{K} \left[ \sum_{i=1}^{N} M_{i\nu} \right] \frac{\sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} M_{i\nu} M_{j\nu} D_{ij}}{\sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} M_{i\nu} M_{j\nu}}, \tag{4}$$

which weights average cluster homogeneities proportional to the cluster size. The total cost thus corresponds to the sum of the average dissimilarities between pairs of objects in the same cluster.

To minimize the cost function in (4) we apply an annealing technique. Optimization methods based on annealing treat the unknown Boolean variables as random variables, introduce a scale parameter $T$, often called the computational temperature, and calculate equilibrium Gibbs averages, e.g., of assignments $M_{i\nu}$. This is achieved either by Monte Carlo sampling or (at least approximately) by analytical methods. The temperature $T$ is gradually lowered and for $T \approx 0$ an admissible solution for the combinatorial optimization problem is found.

Denote by $\mathrm{s}_i^{\nu}(\mathbf{M})$ the matrix obtained by replacing the $i$–th row of $\mathbf{M}$ with the unit vector $\vec{e}_{\,\nu}$ and let $g_{i\nu} = \mathcal{H}(\mathrm{s}_i^{\nu})$. Then an efficient Monte Carlo algorithm is defined by the Gibbs–Sampler, which samples from the conditional probability spanned by site $\vec{x}_i$ for fixed assignments of sites $\vec{x}_j\,, j \neq i$:

$$\mathbf{P}\left(\mathrm{s}_i^{\nu}(\mathbf{M})\right) = \frac{\exp\left(-g_{i\nu}/T\right)}{\sum_{\alpha} \exp\left(-g_{i\alpha}/T\right)}. \tag{5}$$

Note that the calculation of $\mathbf{P}\left(\mathrm{s}_i^{\nu}(\mathbf{M})\right)$ is invariant with respect to additive shifts of the partial costs $g_{i\alpha}$. This can be used to derive an efficient formula for the Gibbs weights of (4). Using the abbreviations $a_{i\nu} = \sum_{k \neq i} M_{k\nu}$, $b_{i\nu} = 2 \sum_{k \in \mathcal{N}_i} M_{k\nu} D_{ik}$, $c_{i\nu} = \sum_{k \neq i} \sum_{l \in \mathcal{N}_k, l \neq i} M_{k\nu} M_{l\nu} D_{kl}$, $o_{i\nu} = 2 \sum_{k \in \mathcal{N}_i} M_{k\nu}$ and $n_{i\nu} = \sum_{k \neq i} \sum_{l \in \mathcal{N}_k, l \neq i} M_{k\nu} M_{l\nu}$ this yields

$$g_{i\nu} = \frac{(a_{i\nu}+1) b_{i\nu} + c_{i\nu}}{n_{i\nu} + o_{i\nu}} - \frac{o_{i\nu} a_{i\nu} c_{i\nu}}{n_{i\nu}(n_{i\nu} + o_{i\nu})}. \tag{6}$$

By the fundamental relationship

$$\langle M_{i\nu} \rangle_Q = \frac{\exp\left(-\langle g_{i\nu} \rangle_Q / T\right)}{\sum_{\alpha} \exp\left(-\langle g_{i\alpha} \rangle_Q / T\right)}, \tag{7}$$

which is valid for factorizing distributions $Q$ minimizing the KL–divergence to $\mathbf{P}$, an even more efficient approximative, deterministic and convergent algorithm with global optimization properties is obtained known as *deterministic annealing*. For the details and a convergence proof, the reader is refered to [4].

## 5 Results

To empirically evaluate the performance of a dissimilarity measure relative to a given set of textures, we have adopted the performance measure proposed in [6] to achieve comparable results. Assume a database of size $m \cdot n$ containing $m$

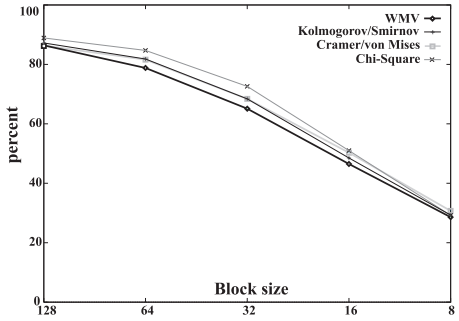


**Figure 4.** Average percentage of correctly identified blocks according to the performance measure proposed by Manjunath and Ma (see text). The results for the Jeffrey–divergence are almost identical to $\chi^2$ and therefore omitted.

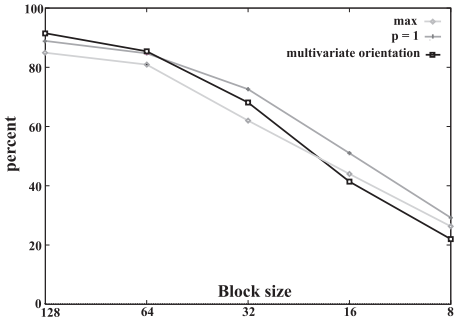


**Figure 5.** Average percentage of correctly identified blocks for $\chi^2$ using different channel integration rules.

samples of each of $n$ textures. For each entry $i$ the distances $d(i,j)$ to all other samples are computed and sorted in ascending order. In the ideal case the first $(m-1)$ textures are of the same type as $i$. The performance is defined as the average percentage of textures of the same type within the top $(m-1)$ matches. It is applied to a database containing 16 random samples for each of 40 Brodatz–like reference textures. For all dissimilarity measures the same Gabor features were used. In Fig. 4 the performance is depicted for the distance measures presented in Sect. 3 and different image (block) sizes. Two important observations are in place:

- The quality for all measures drastically deteriorates for smaller images.
- The statistical similarity measures perform uniformly better than the parametric WMV–measure. The empirical Jeffrey–divergence and the $\chi^2$–test do better than the Cramer/von Mises measure and the KS–statistic.

The first fact was expected, as it becomes even visually more and more difficult to identify small texture patches. WMV implicitly relies on an invalid Gaussian assumption, explaining the inferior quality of the measure, as illustrated

**Figure 6.** Image annotation: The image blocks most similar to the two test blocks, marked by arrows, are depicted by white and gray boxes.

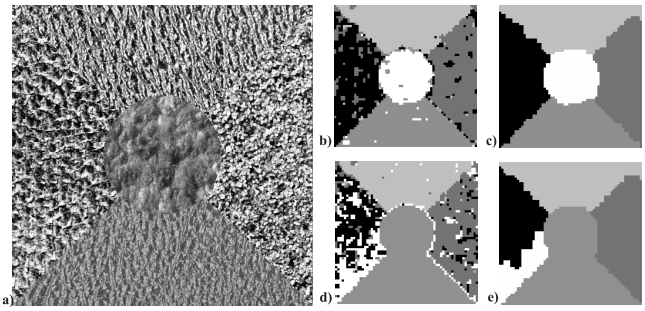


**Figure 7.** Typical segmentation results with $K = 5$: (a) Randomly generated image. (b) Segmentation based on $\chi^2$. (c) Segmentation after post–processing. (d) Segmentation based on WMV. (e) Segmentation after post–processing.

by two example queries in Fig. 1, 2. The measure performs significantly worse than $\chi^2$. We like to stress that the WMV–measure has outperformed several feature–based methods including multiresolution autoregressive models and the tree–structured wavelet transform [6]. The inferior performance of the KS–distance and the Cramer/von Mises measure is explained by the fact, that samples of a not completely uniform texture regularly exhibit a shift in the feature distribution. This results in high dissimilarity values for measures relying on the PDFs.

The quality of a measure depends on the rule to integrate the clues of different feature channels. In Sect. 3 we proposed a family of rules depending on a parameter $p$. As shown in Fig. 5 for the $\chi^2$ distance adding information from the different channels is superior to the max–rule. Alternatively, the Jeffrey–divergence and $\chi^2$ can be applied to multivariate density estimates. This yields superior results for large sample sizes, but suffers from the difficulty to efficiently estimate a multivariate density for few samples. As

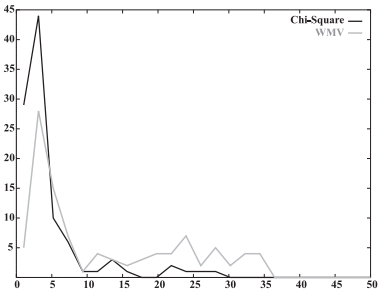


**Figure 8.** Empirical densities of the percentage of misclassified blocks over 100 random images using identical segmentation algorithms.

an example, Fig. 5 shows the results for $\chi^2$ applied to an estimate of the multivariate density of four orientations for a given scale. The univariate $\chi^2$ statistic with $p = 1$ was selected for the subsequent experiments, due to excellent performance and faster computational evaluation compared to the Jeffrey–divergence.

In Fig. 3 three example retrievals for the $\chi^2$–statistic using a large collection of images taken from the VisTex database are shown. The retrieved images are even for the street scene visually similar. Another application is image annotation, where for a given labeled image block similar image regions should be identically labeled in a semi–automatic fashion to speed the annotating process [8]. Figure 6 depicts two test regions and the corresponding most similar regions for an SAR image of Orange County.

From the collection of 40 reference textures we constructed a database containing 100 random mixtures, each of 512x512 pixels and containing five textures (as depicted in Fig. 7(a)). For each image a sub–grid of 64x64 sites and a window size of 16x16 pixels was selected. A typical segmentation example is shown in Fig. 7. All databases and additional examples are available via World Wide Web (WWW). To remove the speckle like noise a simple post–processing step was applied. Note that with the same segmentation algorithm, $\chi^2$ yields a significantly better segmentation than the WMV–measure. Further evidence is given in Fig. 8 which shows the histogram of misclassified blocks with respect to ground truth. The median error rate of 2.65% (7.12% before post–processing) is remarkably good compared to 7.1% (15.7% before post–processing) for the WMV–measure. For $\chi^2$ the essential structure of the image is detected in almost all cases. A typical application for texture segmentation, e.g., in autonomous robotics, are indoor and outdoor images which contain textured objects. An example image of an office environment is presented in Fig. 9. The achieved segmentation is both visually and semantically satisfying. Untextured parts of the image are grouped together irrespectively of there absolute luminance value as expected and the discrimination of the remaining three textures we found highly convincing.

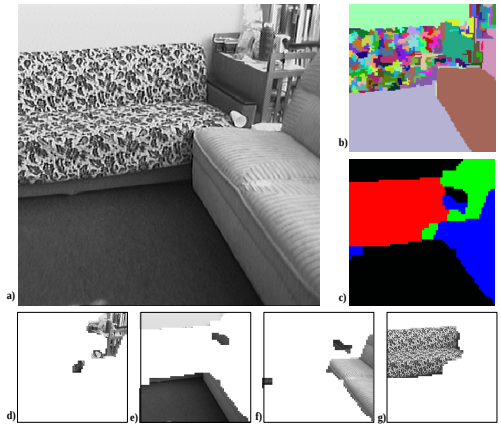


**Figure 9.** (a) Indoor image of a typical office environment containing an old–style sofa. (b) Contrast based image segmentation with a region merging algorithm, (c) a texture segmentation with $K = 4$. The image partitioning is visualized in (d)–(g).

## 6 Summary and Conclusion

We proposed a novel approach for defining similarity measure for textures based on statistical tests to compare the empirical distributions of Gabor coefficients. The major drawback of all feature–based methods, namely the need to specify a suitable metric in parameter space, is hence avoided. An efficient segmentation algorithm operating on the same pairwise similarity values has been presented. The advantages of the unifying framework have been demonstrated by a benchmark on Brodatz–like micro–textures and on real–word images.

## References

[1] M. Flickner et. al. Query by image and video content: The QBIC system. *IEEE Computer*, pages 23–32, Sept. 1995.

[2] D. Geman, S. Geman, C. Graffigne, and P. Dong. Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):609–628, July 1990.

[3] G. Gimel'farb and A. Jain. On retrieving textured images from image database. *Pattern Recognition*, 29(9):1461–1483, 1996.

[4] T. Hofmann, J. Puzicha, and J. Buhmann. A deterministic annealing framework for textured image segmentation. Technical Report IAI-TR-96-2, Institut für Informatik III, 1996.

[5] A. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.

[6] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996.

[7] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[8] R. Picard and T. Minka. Vision texture for annotation. *Multimedia Systems*, 3:3–14, 1995.