ELSEVIER

Pattern Recognition Letters 22 (2001) 563–582

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

Feature normalization and likelihood-based similarity measures for image retrieval

Selim Aksoy *, Robert M. Haralick

Intelligent Systems Laboratory, Department of Electrical Engineering, University of Washington, Seattle, WA 98195-2500, USA

Abstract

Distance measures like the Euclidean distance are used to measure similarity between images in content-based image retrieval. Such geometric measures implicitly assign more weighting to features with large ranges than those with small ranges. This paper discusses the effects of five feature normalization methods on retrieval performance. We also describe two likelihood ratio-based similarity measures that perform significantly better than the commonly used geometric approaches like the L_p metrics. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Feature normalization; Minkowsky metric; Likelihood ratio; Image retrieval; Image similarity

1. Introduction

Image database retrieval has become a very popular research area in recent years (Rui et al., 1999). Initial work on content-based retrieval (Flickner et al., 1993; Pentland et al., 1994; Manjunath and Ma, 1996) focused on using low-level features like color and texture for image representation. After each image is associated with a feature vector, distance measures that compute distances between these feature vectors are used to find similarities between images with the assumption that images that are close to each other in the feature space are also visually similar.

Feature vectors usually exist in a very high dimensional space. Due to this high dimensionality, their parametric characterization is usually not studied, and non-parametric approaches like the nearest neighbor rule are used for retrieval. In geometric similarity measures like the nearest neighbor rule, no assumption is made about the probability distribution of the features and similarity is based on the distances between feature vectors in the feature space. Given this fact, Euclidean (L_2) distance has been the most widely used distance measure (Flickner et al., 1993; Pentland et al., 1994; Li and Castelli, 1997; Smith, 1997). Other popular measures have been the weighted Euclidean distance (Belongie et al., 1998; Rui et al., 1998), the city-block (L_1) distance (Manjunath and Ma, 1996; Smith, 1997), the general Minkowsky L_p distance (Sclaroff et al., 1997) and the Mahalanobis distance (Pentland et al., 1994; Smith, 1997). The L_1 distance was also used under the name "histogram intersection" (Smith, 1997). Berman and Shapiro (1997) used polynomial combinations of predefined distance measures to create new distance measures.

* Corresponding author.
E-mail addresses: aksoy@isl.ee.washington.edu (S. Aksoy), haralick@isl.ee.washington.edu (R.M. Haralick).

PII: S0167-8655(00)00112-4

This paper presents a probabilistic approach for image retrieval. We describe two likelihood-based similarity measures that compute the likelihood of two images being similar or dissimilar, one being the query image and the other one being an image in the database. First, we define two classes, the relevance class and the irrelevance class, and then the likelihood values are derived from a Bayesian classifier. We use two different methods to estimate the conditional probabilities used in the classifier. The first method uses a multivariate Normal assumption and the second one uses independently fitted distributions for each feature. The performances of these two methods are compared to the performances of the commonly used geometric approaches in the form of the L_p metric (e.g., city-block (L_1) and Euclidean (L_2) distances) in ranking the images in the database. We also describe a classification-based criterion to select the best performing p for the L_p metric.

Complex image database retrieval systems use features that are generated by many different feature extraction algorithms with different kinds of sources, and not all of these features have the same range. Popular distance measures, for example the Euclidean distance, implicitly assign more weighting to features with large ranges than those with small ranges. Feature normalization is required to approximately equalize ranges of the features and make them have approximately the same effect in the computation of similarity. In most of the database retrieval literature, the normalization methods were usually not mentioned or only the Normality assumption was used (Manjunath and Ma, 1996; Li and Castelli, 1997; Nastar et al., 1998; Rui et al., 1998). The Mahalanobis distance (Duda and Hart, 1973) also involves normalization in terms of the covariance matrix and produces results related to likelihood when the features are Normally distributed.

This paper discusses five normalization methods: linear scaling to unit range; linear scaling to unit variance; transformation to a Uniform [0,1] random variable; rank normalization; normalization by fitting distributions. The goal is to independently normalize each feature component to the [0,1] range. We investigate the effectiveness of different normalization methods in combination with different similarity measures. Experiments are done on a database of approximately 10,000 images and the retrieval performance is evaluated using average precision and recall computed for a manually groundtruthed data set.

The rest of the paper is organized as follows. First, the features that we use in this study are summarized in Section 2. Then, the feature normalization methods are described in Section 3. Similarity measures for image retrieval are described in Section 4. Experiments and results are discussed in Section 5. Finally, conclusions are given in Section 6.

2. Feature extraction

Textural features that were described in detail by Aksoy and Haralick (1998, 2000b) are used for image representation in this paper. The first set of features are the line-angle-ratio statistics that use a texture histogram computed from the spatial relationships between lines as well as the properties of their surroundings. Spatial relationships are represented by the angles between intersecting line pairs and properties of the surroundings are represented by the ratios of the mean gray levels inside and outside the regions spanned by those angles. The second set of features are the variances of gray level spatial dependencies that use second-order (co-occurrence) statistics of gray levels of pixels in particular spatial relationships. Line-angle-ratio statistics result in a 20-dimensional feature vector and co-occurrence variances result in an 8-dimensional feature vector after the feature selection experiments (Aksoy and Haralick, 2000b).

3. Feature normalization

The following sections describe five normalization procedures. The goal is to independently normalize each feature component to the [0,1] range. A normalization method is preferred over

the others according to the empirical retrieval results that will be presented in Section 5.

3.1. Linear scaling to unit range

Given a lower bound l and an upper bound u for a feature component x,

$$\tilde{x} = \frac{x - l}{u - l} \tag{1}$$

results in $\tilde{x}$ being in the [0,1] range.

3.2. Linear scaling to unit variance

Another normalization procedure is to transform the feature component x to a random variable with zero mean and unit variance as

$$\tilde{x} = \frac{x - \mu}{\sigma}, \tag{2}$$

where μ and σ are the sample mean and the sample standard deviation of that feature, respectively (Jain and Dubes, 1988).

If we assume that each feature is normally distributed, the probability of $\tilde{x}$ being in the [−1,1] range is 68%. An additional shift and rescaling as

$$\tilde{x} = \frac{(x - \mu)/3\sigma + 1}{2} \tag{3}$$

guarantees 99% of $\tilde{x}$ to be in the [0,1] range. We can then truncate the out-of-range components to either 0 or 1.

3.3. Transformation to a Uniform [0,1] random variable

Given a random variable x with cumulative distribution function $F_x(x)$, the random variable $\tilde{x}$ resulting from the transformation $\tilde{x} = F_x(x)$ is uniformly distributed in the [0,1] range (Papoulis, 1991).

3.4. Rank normalization

Given the sample for a feature component for all images as $x_1, \ldots, x_n$, first we find the order statistics $x_{(1)}, \ldots, x_{(n)}$ and then replace each image's feature value by its corresponding normalized rank, as

$$\tilde{x}_i = \frac{\underset{x_1,\ldots,x_n}{\text{rank}}(x_i) - 1}{n - 1}, \tag{4}$$

where x_i is the feature value for the ith image. This procedure uniformly maps all feature values to the [0,1] range. When there are more than one image with the same feature value, for example after quantization, they are assigned the average rank for that value.

3.5. Normalization after fitting distributions

The transformations in Section 3.2 assume that a feature has a Normal (μ, σ^2) distribution. The sample values can be used to find better estimates for the feature distributions. Then, these estimates can be used to find normalization methods based particularly on these distributions.

The following sections describe how to fit Normal, Lognormal, Exponential and Gamma densities to a random sample. We also give the difference distributions because the image similarity measures use feature differences. After estimating the parameters of a distribution, the cut-off value that includes 99% of the feature values is found and the sample values are scaled and truncated so that each feature component have the same range.

Since the original feature values are positive, we use only the positive section of the Normal density after fitting. Lognormal, Exponential and Gamma densities are defined for random variables with only positive values. Other distributions that are commonly encountered in the statistics literature are the Uniform, χ^2 and Weibull (which are special cases of Gamma), Beta (which is defined only for [0,1]) and Cauchy (whose moments do not exist). Although these distributions can also be used by first estimating their parameters and then finding the cut-off values, we will show that the distributions used in this paper can quite generally model features from different feature extraction algorithms.

To measure how well a fitted distribution resembles the sample data (goodness-of-fit), we use the Kolmogorov–Smirnov test statistic (Bury, 1975; Press et al., 1990) which is defined as the maximum value of the absolute difference between the cumulative distribution function estimated from the sample and the one calculated from the fitted distribution. After estimating the parameters for different distributions, we compute the Kolmogorov–Smirnov statistic for each distribution and choose the one with the smallest value as the best fit to our sample.

3.5.1. Fitting a Normal (μ, σ^2) density

Let $x_1, \ldots, x_n \in \mathbb{R}$ be a random sample from a population with density $(1/\sqrt{2\pi}\sigma)\exp(-(x-\mu)^2/2\sigma^2)$, $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$. The likelihood function for the parameters μ and σ^2 is

$$L(\mu, \sigma^2 \,|\, x_1, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum_{i=1}^{n} (x_i - \mu)^2 / 2\sigma^2 \right). \tag{5}$$

After taking its logarithm and equating the partial derivatives to zero, the maximum likelihood estimators (MLEs) of μ and σ^2 can be derived as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2. \tag{6}$$

The cut-off value δ_x that includes 99% of the feature values can be found as

$$P(x \leqslant \delta_x) = P\left(\frac{x - \hat{\mu}}{\hat{\sigma}} \leqslant \frac{\delta_x - \hat{\mu}}{\hat{\sigma}} \right) = 0.99 \Rightarrow \delta_x = \hat{\mu} + 2.4\hat{\sigma}. \tag{7}$$

Let x and y be two iid. random variables with a Normal (μ, σ^2) distribution. Using moment generating functions, we can easily show that their difference $z = x - y$ has a Normal $(0, 2\sigma^2)$ distribution.

3.5.2. Fitting a Lognormal (μ, σ^2) density

Let $x_1, \ldots, x_n \in \mathbb{R}$ be a random sample from a population with density $(1/\sqrt{2\pi}\sigma)(\exp(-(\log x - \mu)^2/2\sigma^2))/x$, $x \geqslant 0$, $-\infty < \mu < \infty$, $\sigma > 0$. The likelihood function for the parameters μ and σ^2 is

$$L(\mu, \sigma^2 \,|\, x_1, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{\exp\left(-\sum_{i=1}^{n} (\log x_i - \mu)^2 / 2\sigma^2 \right)}{\prod_{i=1}^{n} x_i}. \tag{8}$$

The MLEs of μ and σ^2 can be derived as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \log x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (\log x_i - \hat{\mu})^2. \tag{9}$$

In other words, we can take the natural logarithm of each sample point and treat the new data as a sample from a Normal (μ, σ^2) distribution (Casella and Berger, 1990).

The 99% cut-off value δ_x can be found as

$$P(x \leqslant \delta_x) = P(\log x \leqslant \log \delta_x) = P\left(\frac{\log x - \hat{\mu}}{\hat{\sigma}} \leqslant \frac{\log \delta_x - \hat{\mu}}{\hat{\sigma}} \right) = 0.99 \Rightarrow \delta_x = e^{\hat{\mu} + 2.4\hat{\sigma}}. \tag{10}$$

3.5.3. Fitting an Exponential (λ) density

Let $x_1, \ldots, x_n \in \mathbb{R}$ be a random sample from a population with density $(1/\lambda)e^{-x/\lambda}$, $x \geqslant 0$, $\lambda \geqslant 0$. The likelihood function for the parameter λ is

$$L(\lambda \,|\, x_1, \ldots, x_n) = \frac{1}{\lambda^n} \exp\left(-\sum_{i=1}^{n} x_i / \lambda \right). \tag{11}$$

The MLE of λ can be derived as

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{12}$$

The 99% cut-off value δ_x can be found as

$$P(x \leqslant \delta_x) = 1 - e^{-\delta_x/\hat{\lambda}} = 0.99 \Rightarrow \delta_x = -\hat{\lambda}\log 0.01. \tag{13}$$

Let x and y be two iid. random variables with an Exponential(λ) distribution. The distribution of $z = x - y$ can be found as

$$f_z(z) = \frac{1}{2\lambda} e^{-|z|/\lambda}, \quad -\infty < z < \infty. \tag{14}$$

It is called the Double Exponential(λ) distribution and similar to the previous case, the MLE of λ can be derived as

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} |z_i|. \tag{15}$$

3.5.4. Fitting a Gamma(α, β) density

Let $x_1, \ldots, x_n \in \mathbb{R}$ be a random sample from a population with density $(1/\Gamma(\alpha)\beta^{\alpha})x^{\alpha-1} e^{-x/\beta}$, $x \geqslant 0$, $\alpha, \beta \geqslant 0$. Since closed forms for the MLEs of the parameters α and β do not exist, [1] we use the method of moments (MOM) estimators (Casella and Berger, 1990). After equating the first two sample moments to the first two population moments, the MOM estimators for α and β can be derived as

$$\hat{\alpha} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} = \frac{\overline{X}^2}{S^2}, \tag{16}$$

$$\hat{\beta} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)} = \frac{S^2}{\overline{X}}, \tag{17}$$

where $\overline{X}$ and S^2 are the sample mean and the sample variance, respectively.

It can be shown (Casella and Berger, 1990) that when $x \sim \text{Gamma}(\alpha, \beta)$ with an integer α, $P(x \leqslant \delta_x) = P(y \geqslant \alpha)$, where $y \sim \text{Poisson}(\delta_x/\beta)$. Then the 99% cut-off value δ_x can be found as

$$\begin{aligned} P(x \leqslant \delta_x) &= \sum_{y=\hat{\alpha}}^{\infty} e^{-\delta_x/\hat{\beta}} \frac{(\delta_x/\hat{\beta})^y}{y!} \\ &= 1 - \sum_{y=0}^{\hat{\alpha}-1} e^{-\delta_x/\hat{\beta}} \frac{(\delta_x/\hat{\beta})^y}{y!} = 0.99 \\ \Rightarrow \sum_{y=0}^{\hat{\alpha}-1} e^{-\delta_x/\hat{\beta}} &\frac{(\delta_x/\hat{\beta})^y}{y!} = 0.01. \end{aligned} \tag{18}$$

Johnson et al. (1994) represents Eq. (18) as

$$P(x \leqslant \delta_x) = e^{-\delta_x/\hat{\beta}} \sum_{j=0}^{\infty} \frac{(\delta_x/\hat{\beta})^{\hat{\alpha}+j}}{\Gamma(\hat{\alpha}+j+1)}. \tag{19}$$

Another way to find δ_x is to use the Incomplete Gamma function (Abramowitz and Stegun, 1972, p. 260; Press et al., 1990, Section 6.2) as

$$P(x \leqslant \delta_x) = I_{\delta_x/\hat{\beta}}(\hat{\alpha}). \tag{20}$$

Note that unlike Eq. (18), $\hat{\alpha}$ does not have to be an integer in Eq. (20).

Let x and y be two iid. random variables with a Gamma(α, β) distribution. The distribution of $z = x - y$ can be found as (Springer, 1979, p. 356)

$$\begin{aligned} f_z(z) &= \frac{z^{(2\alpha-1)/2}}{(2\beta)^{(2\alpha-1)/2}} \frac{1}{\pi^{1/2}} \frac{1}{\beta\Gamma(\alpha)} K_{\alpha-1/2}(z/\beta), \\ &-\infty < z < \infty, \end{aligned} \tag{21}$$

where $K_m(u)$ is the modified Bessel function of the second kind of order m ($m \geqslant 0$, integer) (Springer, 1979, p. 419; Press et al., 1990, Section 6.6).

Histograms and fitted distributions for some of the 28 features are given in Fig. 1. After comparing the Kolmogorov–Smirnov test statistics as the goodness-of-fits, the line-angle-ratio features were decided to be modeled by Exponential densities and the co-occurrence features were decided to be modeled by Normal densities. Histograms of the normalized features are given in Figs. 2 and 3. Histograms of the differences of normalized features are given in Figs. 4 and 5.

Some example feature histograms and fitted distributions from 60 Gabor features (Manjunath and Ma, 1996), 4 QBIC features (Flickner et al., 1993) and 36 moments features (Cheikh et al., 1999) are also given in Fig. 6. This shows that many features from different feature extraction

[1] MLEs of Gamma parameters can be derived in terms of the "Digamma" function and can be computed numerically (Bury, 1975; Press et al., 1990).

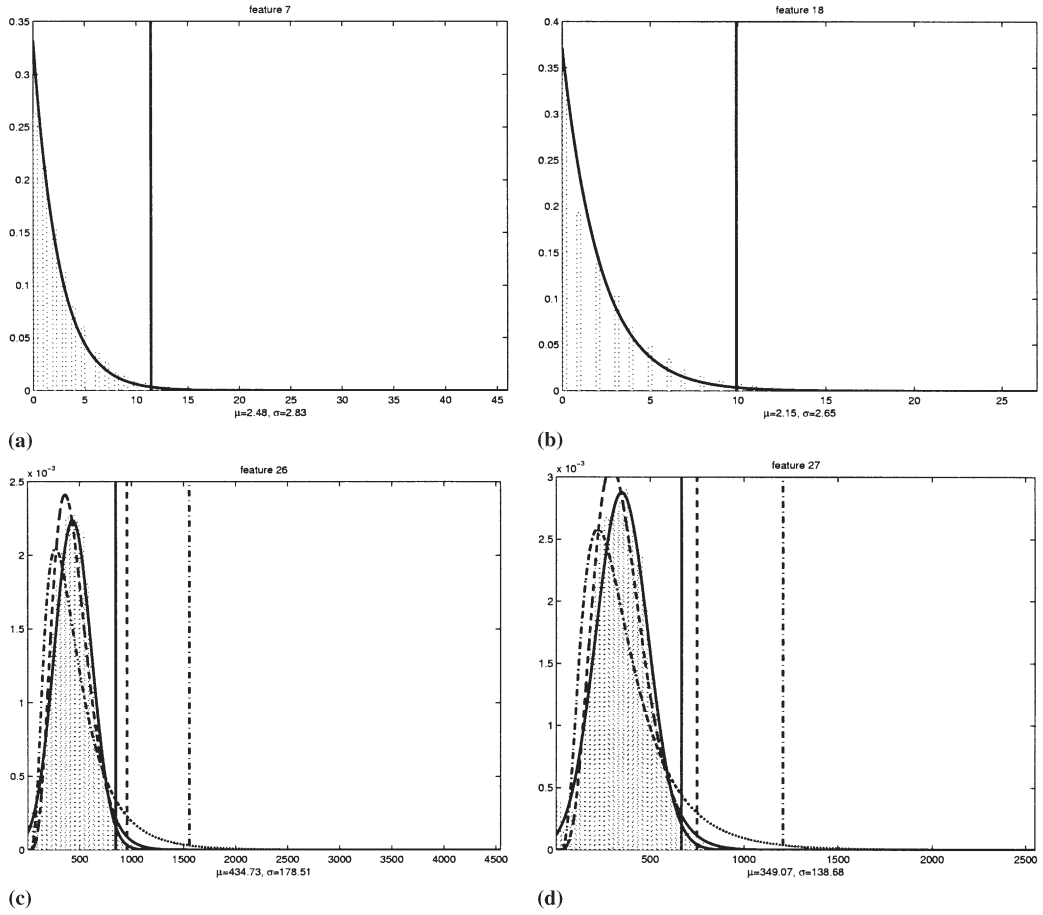

Fig. 1. Feature histograms and fitted distributions for example features. An Exponential model (solid line) is used for the line-angle-ratio features and Normal (solid line), Lognormal (dash–dot line) and Gamma (dashed line) models are used for the co-occurrence features. The vertical lines show the 99% cut-off point for each distribution. (a) Line-angle-ratio (best fit: Exponential); (b) line-angle-ratio (best fit: Exponential); (c) co-occurrence (best fit: Normal); (d) co-occurrence (best fit: Normal).

algorithms can be modeled by the distributions that we presented in Section 3.5.

4. Similarity measures

After computing and normalizing the feature vectors for all images in the database, given a query image, we have to decide which images in the database are relevant to it and we have to retrieve the most relevant ones as the result of the query. A similarity measure for content-based retrieval should be efficient enough to match similar images as well as being able to discriminate dissimilar ones. In this section, we describe two different types of decision methods: likelihood-based probabilistic methods; the nearest neighbor rule with an L_p metric.

4.1. Likelihood-based similarity measures

In our previous work (Aksoy and Haralick, 2000b), we used a two-class pattern classification approach for feature selection. We defined two classes, the relevance class $\mathscr{A}$ and the irrelevance class $\mathscr{B}$, in order to classify image pairs as similar

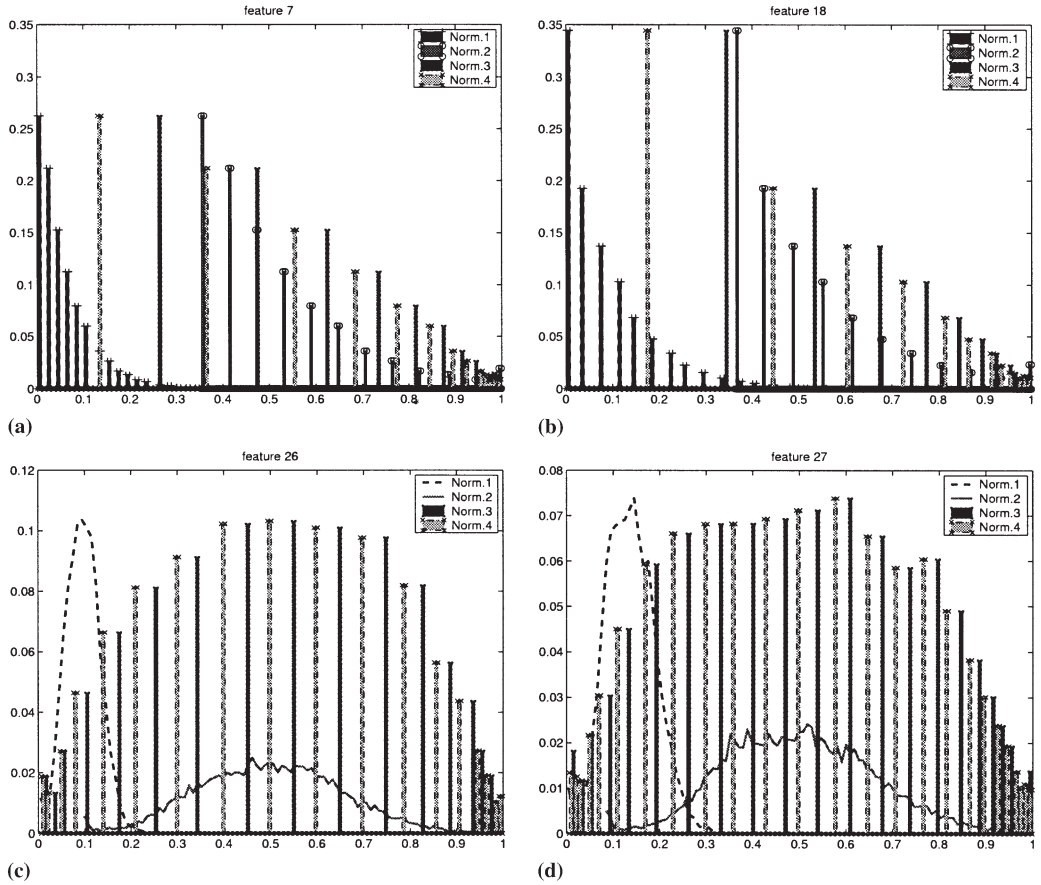

Fig. 2. Normalized feature histograms for the example features in Fig. 1. Numbers in the legends correspond to the normalization methods as follows. Norm.1: linear scaling to unit range; Norm.2: linear scaling to unit variance; Norm.3: transformation to a Uniform[0,1] random variable; Norm.4: rank normalization; Norm.5.1: fitting a Normal density; Norm.5.2: fitting a Lognormal density; Norm.5.3: fitting an Exponential density; Norm.5.4: fitting a Gamma density.

or dissimilar. A Bayesian classifier can be used for this purpose as follows. Given two images with feature vectors x and y, and their feature difference vector $d = x - y$, $x, y, d \in \mathbb{R}^q$ with q being the size of a feature vector, the a posteriori probability that they are relevant is

$$P(\mathscr{A}\,|\,d) = P(d\,|\,\mathscr{A})P(\mathscr{A})/P(d) \tag{22}$$

and the a posteriori probability that they are irrelevant is

$$P(\mathscr{B}\,|\,d) = P(d\,|\,\mathscr{B})P(\mathscr{B})/P(d). \tag{23}$$

Assuming that these two classes are equally likely, the likelihood ratio is defined as

$$r(d) = \frac{P(d\,|\,\mathscr{A})}{P(d\,|\,\mathscr{B})}. \tag{24}$$

In the following sections, we describe two methods to estimate the conditional probabilities $P(d\,|\,\mathscr{A})$ and $P(d\,|\,\mathscr{B})$. The class-conditional densities are represented in terms of feature difference vectors because similarity between images is assumed to be based on the closeness of their feature values, i.e. similar images have similar feature values (therefore, a difference vector with zero mean and a small variance) and dissimilar images have relatively different feature values

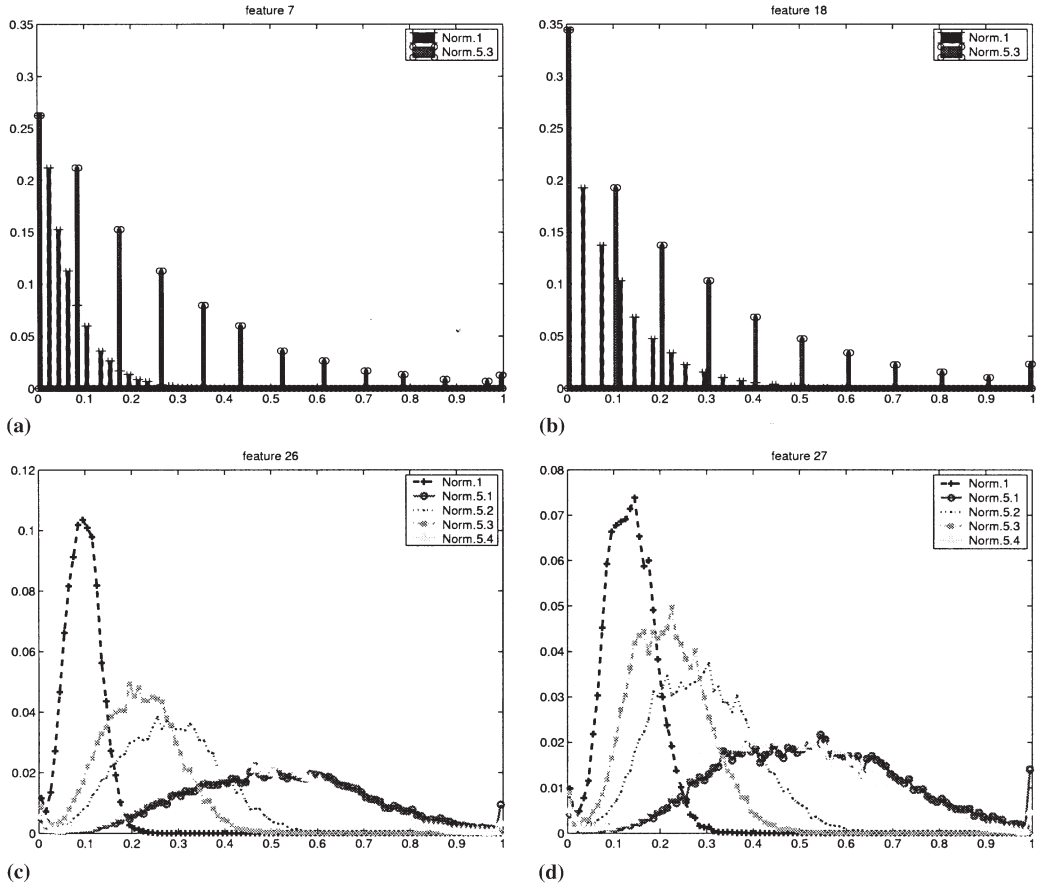

Fig. 3. Normalized feature histograms for the example features in Fig. 1 (cont.). Numbers in the legends are described in Fig. 2.

(a difference vector with a non-zero mean and a large variance).

4.1.1. Multivariate Normal assumption

We assume that the feature differences for the relevance class have a multivariate Normal density with mean $\mu_{\mathscr{A}}$ and covariance matrix $\Sigma_{\mathscr{A}}$,

$$f(d\,|\,\mu_{\mathscr{A}},\Sigma_{\mathscr{A}}) = \frac{1}{(2\pi)^{q/2}|\Sigma_{\mathscr{A}}|^{1/2}} \times \exp\Big(-(d-\mu_{\mathscr{A}})'\Sigma_{\mathscr{A}}^{-1}(d-\mu_{\mathscr{A}})/2\Big). \tag{25}$$

Similarly, the feature differences for the irrelevance class are assumed to have a multivariate Normal density with mean $\mu_{\mathscr{B}}$ and covariance matrix $\Sigma_{\mathscr{B}}$,

$$f(d\,|\,\mu_{\mathscr{B}},\Sigma_{\mathscr{B}}) = \frac{1}{(2\pi)^{q/2}|\Sigma_{\mathscr{B}}|^{1/2}} \times \exp\Big(-(d-\mu_{\mathscr{B}})'\Sigma_{\mathscr{B}}^{-1}(d-\mu_{\mathscr{B}})/2\Big). \tag{26}$$

The likelihood ratio in Eq. (24) is given as

$$r(d) = \frac{f(d\,|\,\mu_{\mathscr{A}},\Sigma_{\mathscr{A}})}{f(d\,|\,\mu_{\mathscr{B}},\Sigma_{\mathscr{B}})}. \tag{27}$$

Given training feature difference vectors $d_1,\ldots,d_n \in \mathbb{R}^q$, $\mu_{\mathscr{A}}$, $\Sigma_{\mathscr{A}}$, $\mu_{\mathscr{B}}$ and $\Sigma_{\mathscr{B}}$ are estimated using the multivariate versions of the MLEs given in Section 3.5.1 as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(d_i-\hat{\mu})(d_i-\hat{\mu})'. \tag{28}$$

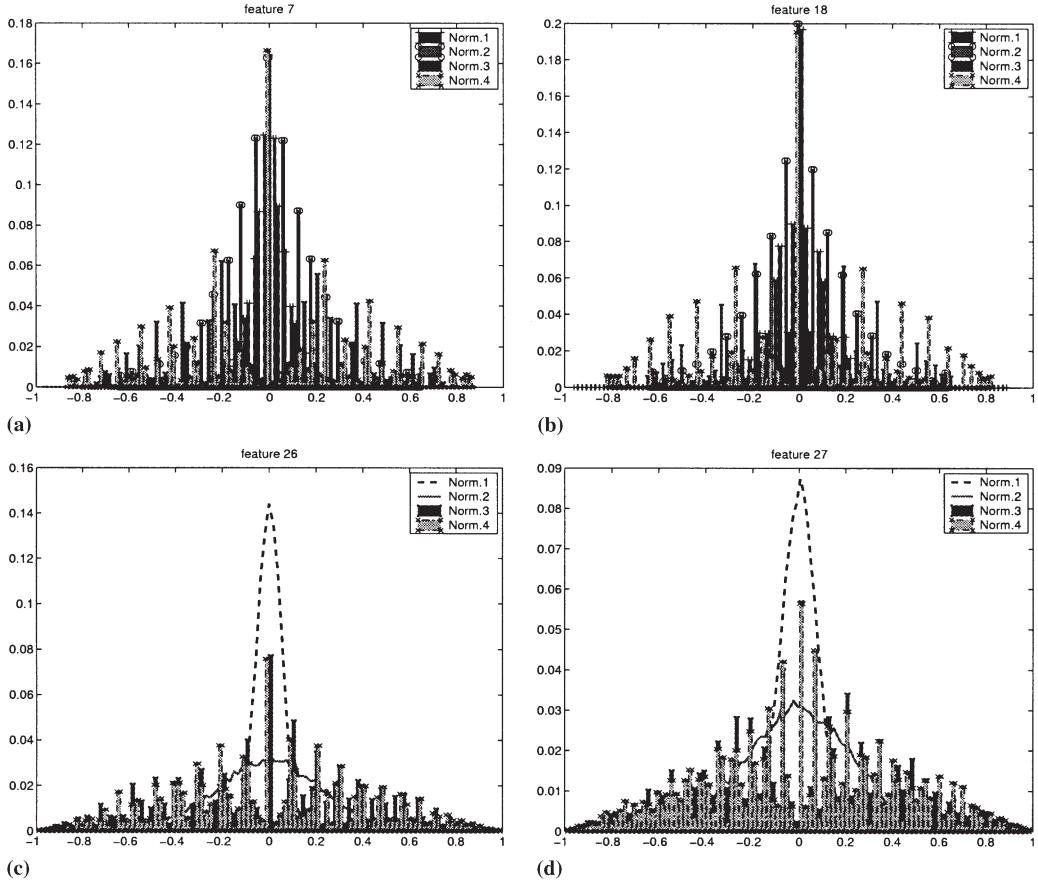

Fig. 4. Histograms of the differences of the normalized features in Fig. 2. Numbers in the legends are described in Fig. 2.

To simplify the computation of the likelihood ratio in Eq. (27), we take its logarithm, eliminate some constants, and use

$$r(d) = (d-\mu_{\mathscr{A}})'\Sigma_{\mathscr{A}}^{-1}(d-\mu_{\mathscr{A}}) - (d-\mu_{\mathscr{B}})'\Sigma_{\mathscr{B}}^{-1}(d-\mu_{\mathscr{B}}) \tag{29}$$

to rank the database images in ascending order of these values which corresponds to a descending order of similarity. This ranking is equivalent to ranking in descending order using the likelihood values in Eq. (27).

4.1.2. Independently fitted distributions

We also use the fitted distributions to compute the likelihood values. Using the Double Exponential model in Section 3.5.3 for the 20 line-angle-ratio feature differences and the Normal model in Section 3.5.1 for the 8 co-occurrence feature differences independently for each feature component, the joint density for the relevance class is given as

$$f(d\,|\,\lambda_{\mathscr{A}1},\ldots,\lambda_{\mathscr{A}20},\mu_{\mathscr{A}21},\ldots,\mu_{\mathscr{A}28},\sigma^2_{\mathscr{A}21},\ldots,\sigma^2_{\mathscr{A}28}) = \prod_{i=1}^{20}\frac{1}{2\lambda_{\mathscr{A}i}}e^{-|d_i|/\lambda_{\mathscr{A}i}}\prod_{i=21}^{28}\frac{1}{\sqrt{2\pi\sigma^2_{\mathscr{A}i}}}e^{-(d_i-\mu_{\mathscr{A}i})^2/2\sigma^2_{\mathscr{A}i}} \tag{30}$$

and the joint density for the irrelevance class is given as

$$f(d\,|\,\lambda_{\mathscr{B}1},\ldots,\lambda_{\mathscr{B}20},\mu_{\mathscr{B}21},\ldots,\mu_{\mathscr{B}28},\sigma^2_{\mathscr{B}21},\ldots,\sigma^2_{\mathscr{B}28}) = \prod_{i=1}^{20}\frac{1}{2\lambda_{\mathscr{B}i}}e^{-|d_i|/\lambda_{\mathscr{B}i}}\prod_{i=21}^{28}\frac{1}{\sqrt{2\pi\sigma^2_{\mathscr{B}i}}}e^{-(d_i-\mu_{\mathscr{B}i})^2/2\sigma^2_{\mathscr{B}i}}. \tag{31}$$

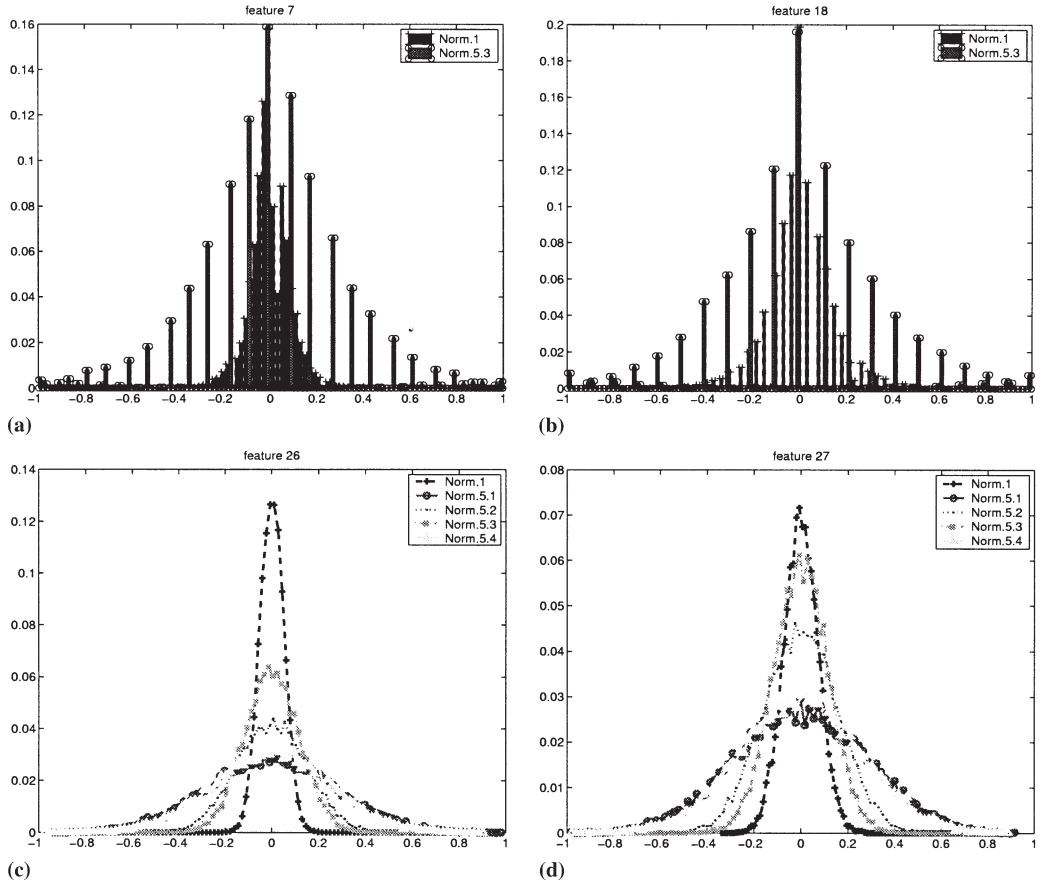

Fig. 5. Histograms of the differences of the normalized features in Fig. 3. Numbers in the legends are described in Fig. 2.

The likelihood ratio in Eq. (24) becomes

$$r(d) = \frac{f(d\,|\,\lambda_{\mathscr{A}1},\ldots,\lambda_{\mathscr{A}20},\mu_{\mathscr{A}21},\ldots,\mu_{\mathscr{A}28},\sigma^2_{\mathscr{A}21},\ldots,\sigma^2_{\mathscr{A}28})}{f(d\,|\,\lambda_{\mathscr{B}1},\ldots,\lambda_{\mathscr{B}20},\mu_{\mathscr{B}21},\ldots,\mu_{\mathscr{B}28},\sigma^2_{\mathscr{B}21},\ldots,\sigma^2_{\mathscr{B}28})}. \tag{32}$$

$\lambda_{\mathscr{A}i},\lambda_{\mathscr{B}i},\mu_{\mathscr{A}i},\mu_{\mathscr{B}i},\sigma^2_{\mathscr{A}i},\sigma^2_{\mathscr{B}i}$ are estimated using the MLEs given in Sections 3.5.1 and 3.5.3. Instead of computing the complete likelihood ratio, we take its logarithm, eliminate some constants, and use

$$r(d) = \sum_{i=1}^{20} |d_i| \left(\frac{1}{\lambda_{\mathscr{A}i}} - \frac{1}{\lambda_{\mathscr{B}i}} \right) + \frac{1}{2} \sum_{i=21}^{28} \left[\frac{(d_i - \mu_{\mathscr{A}i})^2}{\sigma^2_{\mathscr{A}i}} - \frac{(d_i - \mu_{\mathscr{B}i})^2}{\sigma^2_{\mathscr{B}i}} \right] \tag{33}$$

to rank the database images.

4.2. The nearest neighbor rule

In the geometric similarity measures like the nearest neighbor decision rule, each image n in the database is assumed to be represented by its feature vector $y^{(n)}$ in the q-dimensional feature space. Given the feature vector x for the input query, the goal is to find the ys which are the closest neighbors of x according to a distance measure ρ. Then, the k-nearest neighbors of x will be retrieved as the most relevant ones.

The problem of finding the k-nearest neighbors can be formulated as follows. Given the set $Y = \{y^{(n)}\,|\,y^{(n)} \in \mathbb{R}^q, n = 1,\ldots,N\}$ and feature vector $x \in \mathbb{R}^q$, find the set of images $U \subseteq \{1,\ldots,N\}$ such that $\#U = k$ and

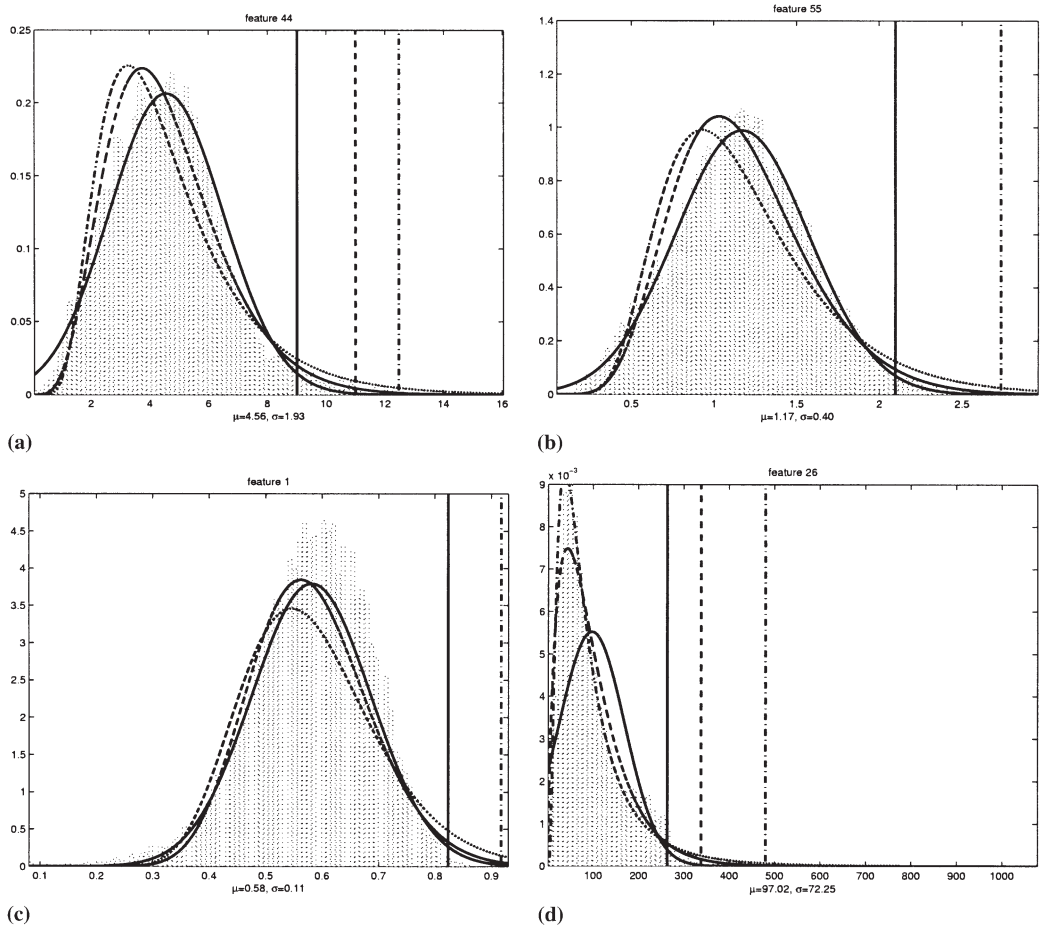

Fig. 6. Feature histograms and fitted distributions for some of the Gabor, QBIC and moments features. The vertical lines show the 99% cut-off points. (a) Gabor (best fit: Gamma); (b) Gabor (best fit: Normal); (c) QBIC (best fit: Normal); (d) moments (best fit: Lognormal).

$$\rho(x, y^{(u)}) \leqslant \rho(x, y^{(v)}), \quad \forall u \in U, v \in \{1, \ldots, N\} \setminus U, \tag{34}$$

where N being the number of images in the database. Then, images in the set U are retrieved as the result of the query.

4.2.1. The L_p metric

As the distance measure, we use the Minkowsky L_p metric (Naylor and Sell, 1982)

$$\rho_p(x, y) = \left(\sum_{i=1}^{q} |x_i - y_i|^p \right)^{1/p}, \tag{35}$$

for $p \geqslant 1$, where $x, y \in \mathbb{R}^q$ and x_i and y_i are the ith components of the feature vectors x and y, respectively. A modified version of the L_p metric as

$$\rho_p(x, y) = \sum_{i=1}^{q} |x_i - y_i|^p \tag{36}$$

is also a metric for $0 < p < 1$. We use the form in Eq. (36) for $p > 0$ to rank the database images since the power $1/p$ in Eq. (35) does not affect the ranks. We will describe how we choose which p to use in the following section.

4.2.2. Choosing p

Commonly used forms of the L_p metric are the city-block (L_1) distance and the Euclidean (L_2) distance. Sclaroff et al. (1997) used L_p metrics with a selection criterion based on the relevance feedback from the user. The best L_p metric for each query was chosen as the one that minimized the average distance between the images labeled as relevant. However, no study of the performance of this selection criterion was presented.

We use a linear classifier to choose the best p value for the L_p metric. Given training sets of feature vector pairs (x, y) for the relevance and irrelevance classes, first, the distances ρ_p are computed as in Eq. (36). Then, from the histograms of ρ_p for the relevance class $\mathscr{A}$ and the irrelevance class $\mathscr{B}$, a threshold θ is selected for classification. This corresponds to a likelihood ratio test where the class-conditional densities are estimated by the histograms.

After the threshold is selected, the classification rule becomes

$$\text{assign } (x,y) \text{ to } \begin{cases} \text{class } \mathscr{A} & \text{if } \rho_p(x,y) < \theta, \\ \text{class } \mathscr{B} & \text{if } \rho_p(x,y) \geqslant \theta. \end{cases} \tag{37}$$

We use a minimum error decision rule with equal priors, i.e. the threshold is the intersecting point of the two histograms. The best p value is then chosen as the one that minimizes the classification error which is 0.5 misdetection + 0.5 false alarm.

5. Experiments and results

5.1. Database population

Our database contains 10,410 256×256 images that came from the Fort Hood Data of the RADIUS Project and also from the LANDSAT and Defense Meteorological Satellite Program (DMSP) Satellites. The RADIUS images consist of visible light aerial images of the Fort Hood area in Texas, USA. The LANDSAT images are from a remote sensing image collection.

Two traditional measures for retrieval performance in the information retrieval literature are precision and recall. Given a particular number of images retrieved, precision is defined as the percentage of retrieved images that are actually relevant and recall is defined as the percentage of relevant images that are retrieved. For these tests,

Table 1
Average precision when 18 images are retrieved[a]

p \ Method	Norm.1	Norm.2	Norm.3	Norm.4	Norm.5.1	Norm.5.2	Norm.5.3	Norm.5.4
0.4	0.4615	0.4801	0.5259	0.5189	0.4777	0.4605	0.4467	0.4698
0.5	0.4741	0.4939	0.5404	0.5298	0.4936	0.4706	0.4548	0.4828
0.6	0.4800	0.5018	0.5493	0.5378	0.5008	0.4773	**0.4586**	0.4892
0.7	**0.4840**	0.5115	0.5539	0.5423	0.5033	**0.4792**	0.4582	0.4910
0.8	0.4837	0.5117	0.5562	0.5457	0.5078	0.4778	0.4564	**0.4957**
0.9	0.4830	**0.5132**	0.5599	0.5471	**0.5090**	0.4738	0.4553	0.4941
1.0	0.4818	0.5117	0.5616	0.5457	0.5049	0.4731	0.4552	0.4933
1.1	0.4787	0.5129	0.5626	0.5479	0.5048	0.4749	0.4510	0.4921
1.2	0.4746	0.5115	0.5641	0.5476	0.5032	0.4737	0.4450	0.4880
1.3	0.4677	0.5112	0.5648	0.5476	0.4995	0.4651	0.4369	0.4825
1.4	0.4632	0.5065	0.5661	**0.5482**	0.4973	0.4602	0.4342	0.4803
1.5	0.4601	0.5052	**0.5663**	0.5457	0.4921	0.4537	0.4303	0.4737
1.6	0.4533	0.5033	0.5634	0.5451	0.4868	0.4476	0.4231	0.4692
2.0	0.4326	0.4890	0.5618	0.5369	0.4755	0.4311	0.4088	0.4536

[a] Columns represent different normalization methods. Rows represent different p values. The largest average precision for each normalization method is marked in bold. The normalization methods that are used in the experiments are represented as: Norm.1: linear scaling to unit range; Norm.2: linear scaling to unit variance; Norm.3: transformation to a Uniform[0,1] random variable; Norm.4: rank normalization; Norm.5.1: fitting Exponentials to line-angle-ratio features and fitting Normals to co-occurrence features; Norm.5.2: fitting Exponentials to line-angle-ratio features and fitting Lognormals to co-occurrence features; Norm.5.3: fitting Exponentials to all features; Norm.5.4: fitting Exponentials to line-angle-ratio features and fitting Gammas to co-occurrence features.

we randomly selected 340 images from the total of 10,410 and formed a groundtruth of seven categories; parking lots, roads, residential areas, landscapes, LANDSAT USA, DMSP North Pole and LANDSAT Chernobyl.

The training data for the likelihood-based similarity measures was generated using the protocol described in (Aksoy and Haralick, 1998). This protocol divides an image into sub-images which overlap by at most half their area and records the relationships between them. Since the original images from the Fort Hood Dataset that we use as the training set have a lot of structure, we assume that sub-image pairs that overlap are relevant (training data for the relevance class) and sub-image pairs that do not overlap are usually not relevant (training data for the irrelevance class).

The normalization methods that are used in the experiments in the following sections are indicated in the caption to Table 1. The legends in the following figures refer to the same caption.

5.2. Choosing p for the L_p metric

The p values that were used in the experiments below were chosen using the approach described

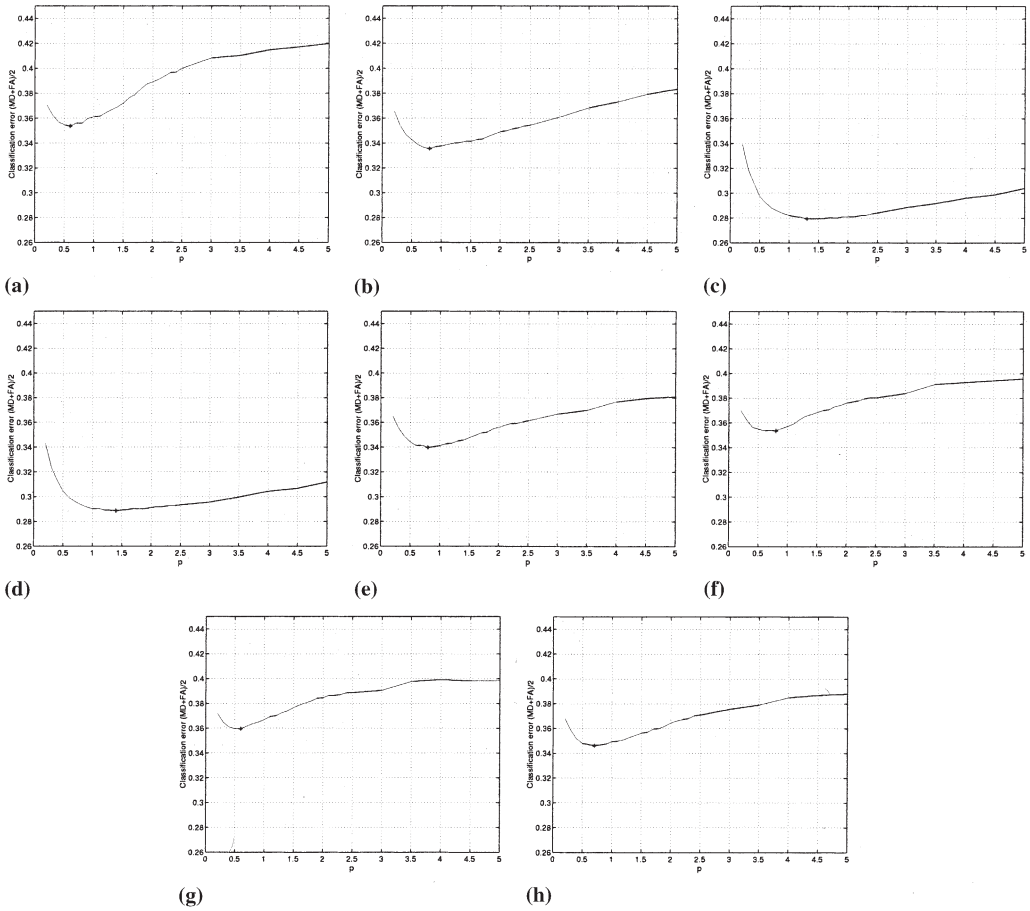

Fig. 7. Classification error vs. p for different normalization methods. The best p value is marked for each method. (a) Norm.1 (best $p = 0.6$); (b) Norm.2 (best $p = 0.8$); (c) Norm.3 (best $p = 1.3$); (d) Norm.4 (best $p = 1.4$); (e) Norm.5.1 (best $p = 0.8$); (f) Norm.5.2 (best $p = 0.8$); (g) Norm.5.3 (best $p = 0.6$); (h) Norm.5.4 (best $p = 0.7$).

in Section 4.2.2. For each normalization method, we computed the classification error for p in the range [0.2,5]. The results are given in Fig. 7. We also computed the average precision for all normalization methods for p in the range [0.4,2] as given in Table 1. The values of p that resulted in the smallest classification error and the largest average precision were consistent. Therefore, the classification scheme presented in Section 4.2.2 was an effective way of deciding which p to use. The p values that gave the smallest classification error for each normalization method were used in the retrieval experiments of the following section.

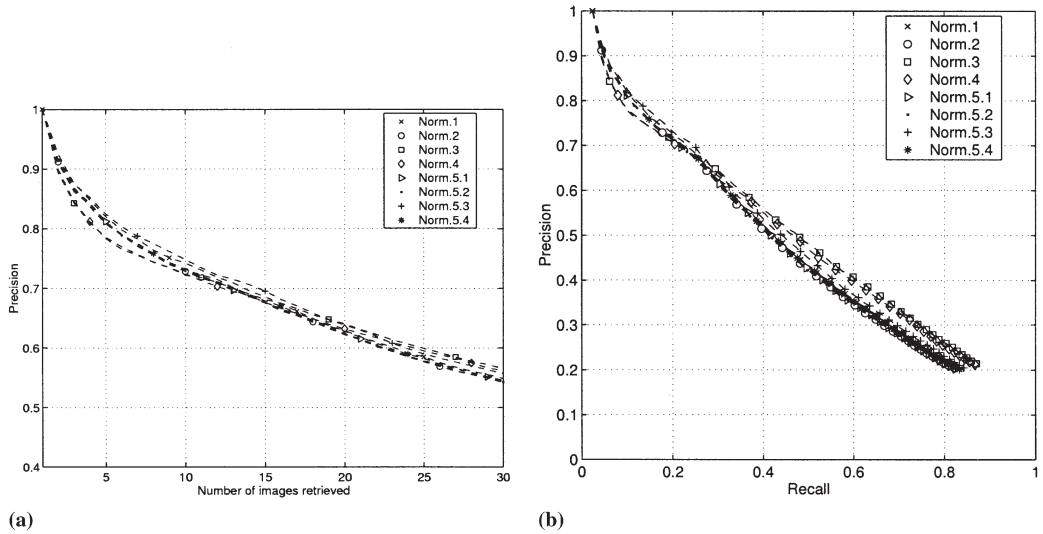

Fig. 8. Retrieval performance for the whole database using the likelihood ratio with the multivariate Normal assumption. (a) Precision vs. number of images retrieved; (b) precision vs. recall.

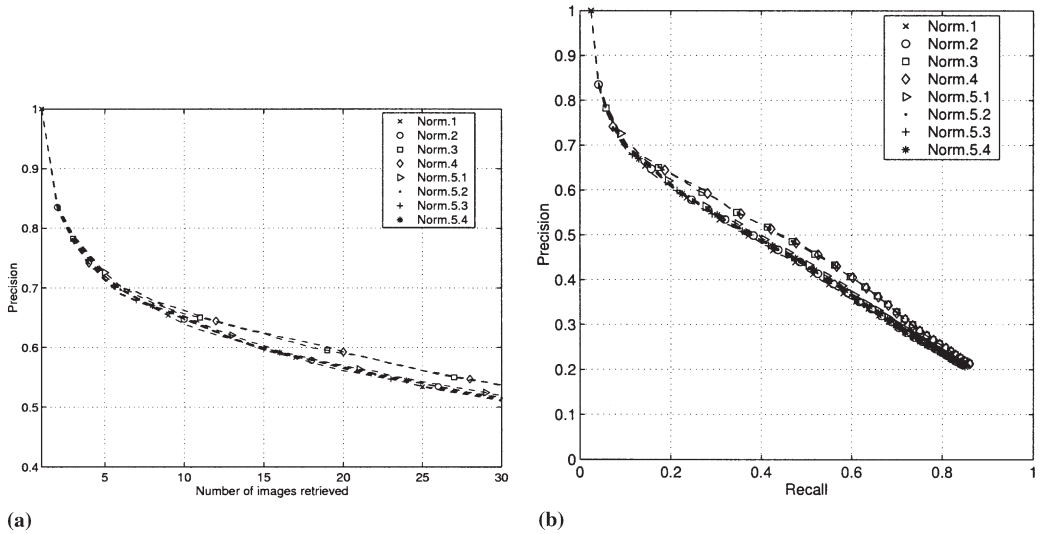

Fig. 9. Retrieval performance for the whole database using the likelihood ratio with the fitted distributions. (a) Precision vs. number of images retrieved; (b) precision vs. recall.

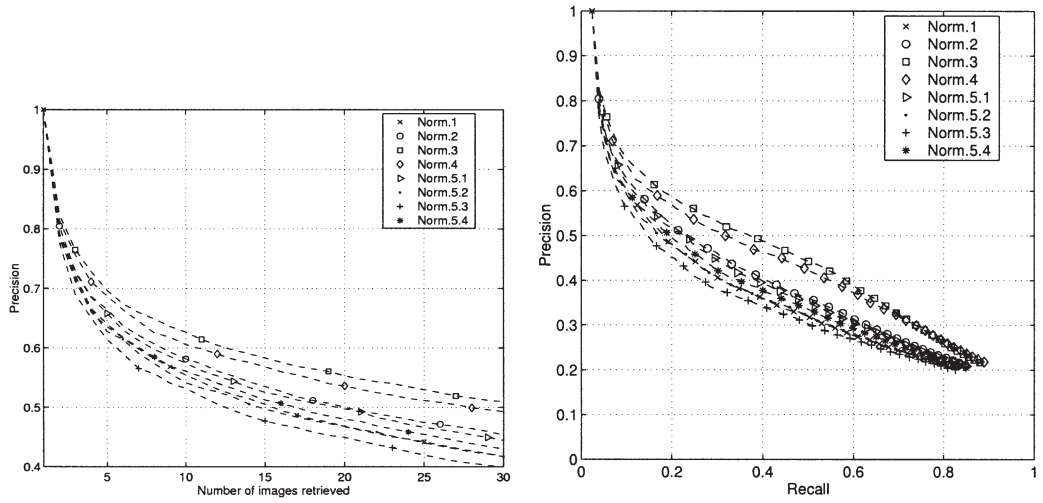

Fig. 10. Retrieval performance for the whole database using the L_p metric. (a) Precision vs. number of images retrieved; (b) precision vs. recall.

5.3. Retrieval performance

Retrieval results, in terms of precision and recall averaged over the groundtruth images, for the likelihood ratio with multivariate Normal assumption, the likelihood ratio with fitted distributions and the L_p metric with different normalization methods are given in Figs. 8–10, respectively. Note that, linear scaling to unit range involves only scaling and translation and it does not have any truncation so it does not change the structures of distributions of the features. Therefore, using this method reflects the effects of using the raw feature distributions while mapping them to the same range. Figs. 11 and 12 show the retrieval performance for the normalization methods separately. Example queries using the same query image but different similarity measures are given in Fig. 13.

5.4. Observations

- Using probabilistic similarity measures always performed better in terms of both precision and recall than the cases where the geometric measures with the L_p metric were used. On the average, the likelihood ratio that used the multivariate Normality assumption performed better than the likelihood ratio that used independent features with fitted Exponential or Normal distributions. The covariance matrix in the correlated multivariate Normal captured more information than using individually better fitted but assumed to be independent distributions.
- Probabilistic measures performed similarly when different normalization methods were used. This shows that these measures are more robust to normalization effects than the geometric measures. The reason for this is that the parameters used in the class-conditional densities (e.g. covariance matrix) were estimated from the normalized features, therefore the likelihood-based methods have an additional (built-in) normalization step.
- The L_p metric performed better for values of p around 1. This is consistent with our earlier experiments where the city-block (L_1) distance performed better than the Euclidean (L_2) distance (Aksoy and Haralick, 2000a). Different normalization methods resulted in different ranges of best performing p values in the classification tests for the L_p metric. Both the smallest classification error and the largest average precision were obtained with

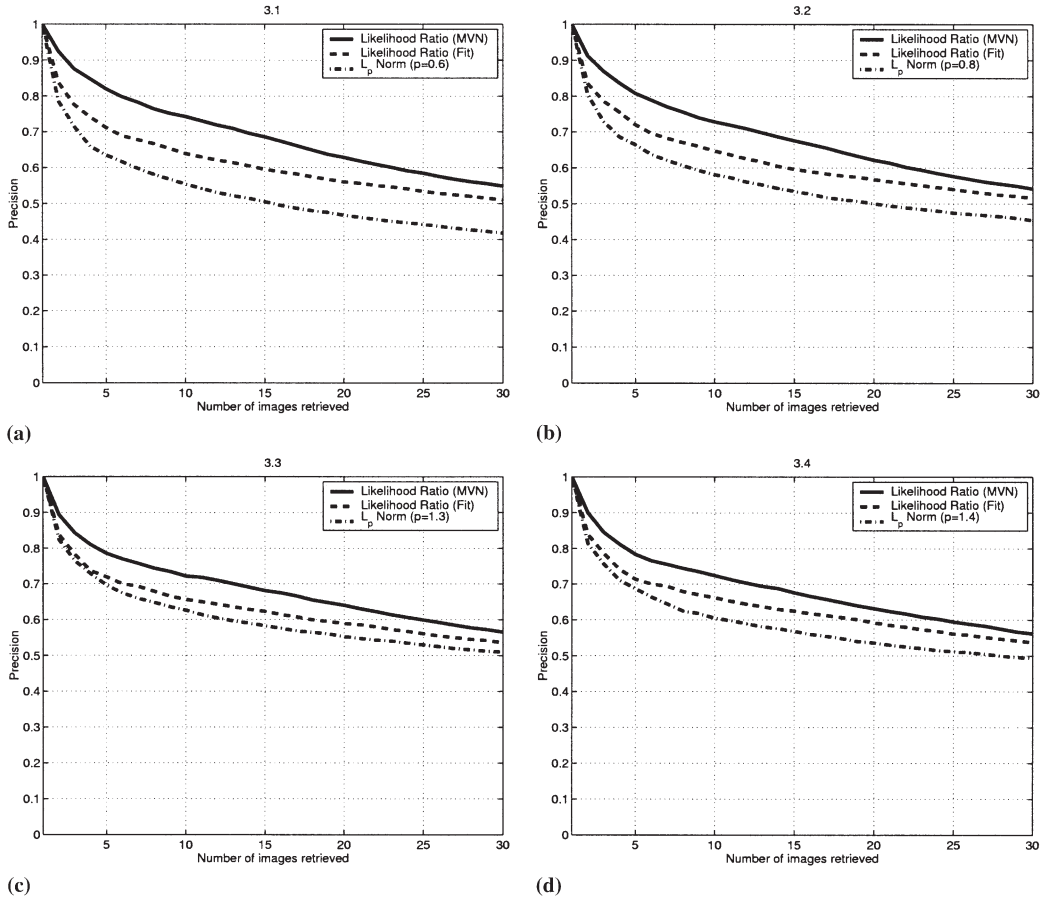

Fig. 11. Precision vs. number of images retrieved for the similarity measures used with different normalization methods. (a) Linear scaling to unit range (Norm.1); (b) linear scaling to unit range (Norm.2); (c) transformation to a Uniform r.v (Norm.3); (d) rank normalization (Norm.4).

normalization methods like transformation using the cumulative distribution function (Norm.3) or the rank normalization (Norm.4), i.e. the methods that tend to make the distribution uniform. These methods also resulted in relatively flat classification error curves around the best performing p values which showed that a larger range of p values were performing similarly well. Therefore, flat minima are indicative of a more robust method. All the other methods had at least 20% worse classification errors and 10% worse precisions. They were also more sensitive to the choice of p and both the classification error and the average precision changed fast with smaller changes in p. Besides, the values of p that resulted in both the smallest classification errors and the largest average precisions were consistent. Therefore, the classification scheme presented in Section 4.2.2 was an effective way of deciding which p to use in the L_p metric.

- The best performing p values for the methods Norm.3 and Norm.4 were around 1.5 whereas smaller p values around 0.7 performed better for other methods. Given the structure of the L_p metric, a few relatively large differences can effect the results significantly for larger p values. On the other hand, smaller p values are less sensitive to large differences. Therefore, smaller p values tend to make a distance more robust to

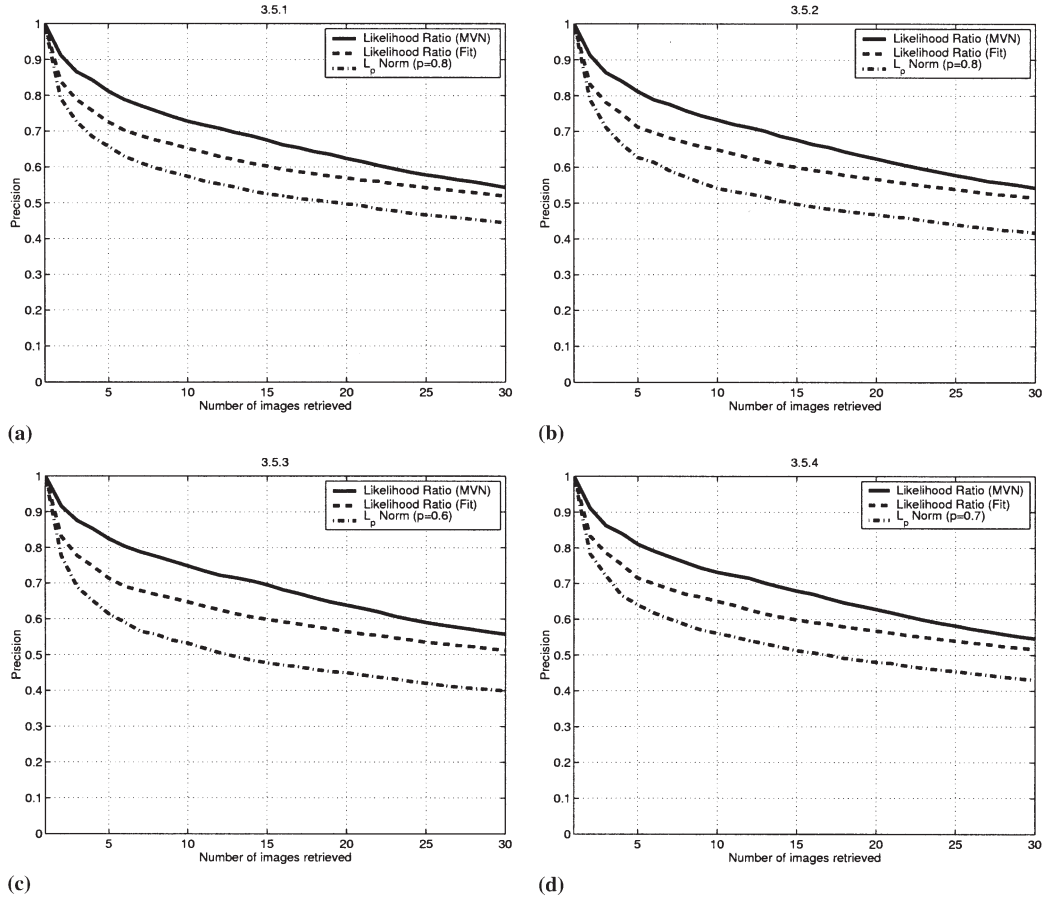

Fig. 12. Precision vs. number of images retrieved for the similarity measures used with different normalization methods (cont.). (a) Fitting Exponentials and Normals (Norm.5.1); (b) fitting Exponentials and Lognormals (Norm.5.2); (c) fitting Exponentials (Norm.5.3); (d) fitting Exponentials and Gammas (Norm.5.4).

large differences. This is consistent with the fact that L_1-regression is more robust than least squares with respect to outliers (Rousseeuw and Leroy, 1987). This shows that the normalization methods other than Norm.3 and Norm.4 resulted in relatively unbalanced feature spaces and smaller p values tried to reduce this effect in the L_p metric.

- Using only scaling to unit range performed worse than most of the other methods. This is consistent with the observation that spreading out the feature values in the [0,1] range as much as possible improved the discrimination capabilities of the L_p metrics.
- Among the methods with fitting distributions, fitting Exponentials to the line-angle-ratio features and fitting Normals to the co-occurrence features performed better than others. We can conclude that studying the distributions of the features and using the results of this study significantly improves the results compared to making only general or arbitrary assumptions.

6. Conclusions

This paper investigated the effects of feature normalization on retrieval performance in an

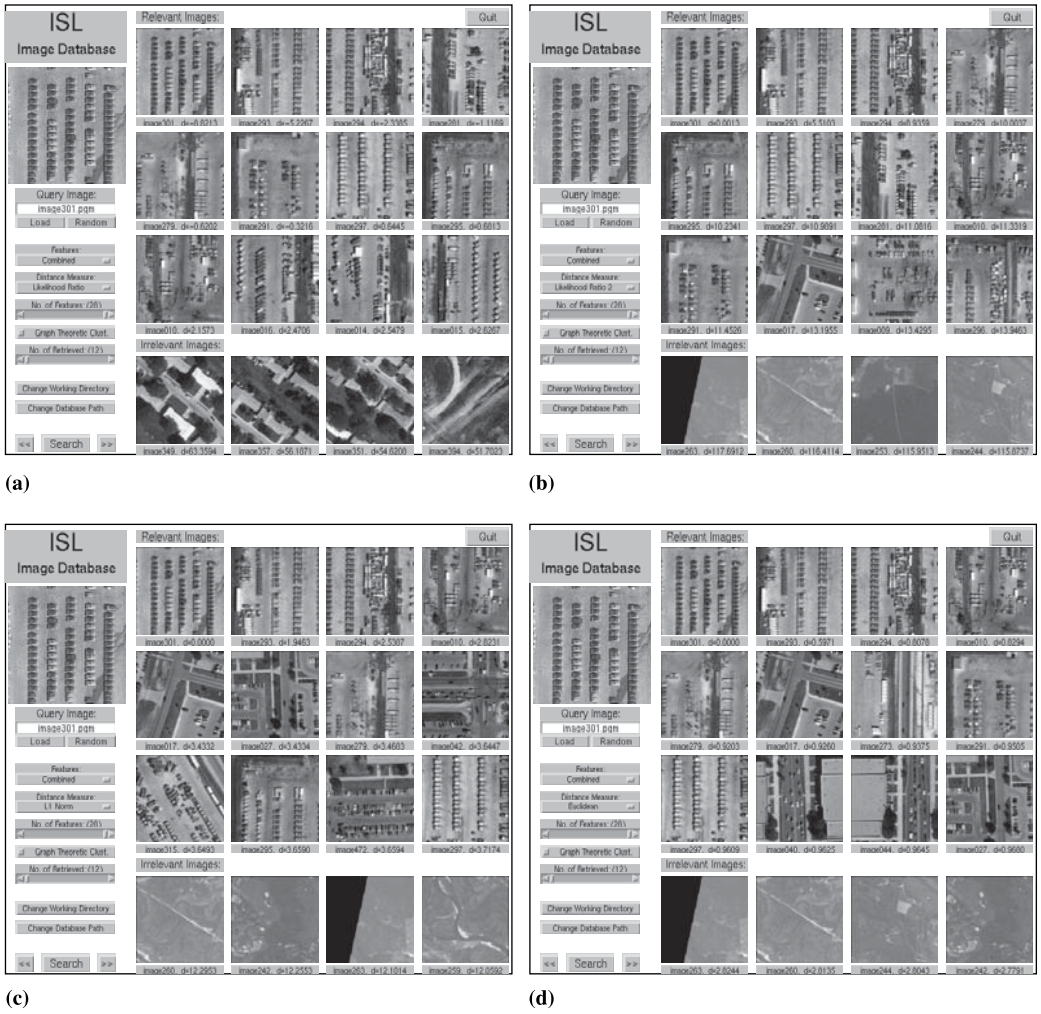

Fig. 13. Retrieval examples using the same parking lot image as query with different similarity measures. The upper left image is the query. Among the retrieved images, first three rows show the 12 most relevant images in descending order of similarity and the last row shows the 4 most irrelevant images in descending order of dissimilarity. Please note that both the order and the number of similar images retrieved with different measures are different. (a) Likelihood ratio (MVN) (12 similar images retrieved); (b) likelihood ratio (fit) (11 similar images retrieved); (c) city-block (L_1) distance (9 similar images retrieved); (d) Euclidean (L_2) distance (7 similar images retrieved).

image database retrieval system. We described five normalization methods: linear scaling to unit range; linear scaling to unit variance; transformation to a Uniform [0,1] random variable; rank normalization; normalization by fitting distributions to independently normalize each feature component to the [0,1] range. We showed that the features were not always Normally distributed as usually assumed, and normalization with respect to a fitted distribution was required. We also showed that many features that were computed by different feature extraction algorithms could be modeled by the methods that we presented, and spreading out the feature values in the [0,1] range as much as possible improved the discrimination capabilities of the similarity measures. The best results were obtained with the normalization methods of transformation using the cumulative

distribution function and rank normalization. The final choice for the normalization method that will be used in a retrieval system will depend on the precision and recall results for the specific data set after applying the methods presented in this paper.

We also described two new probabilistic similarity measures and compared their retrieval performances with those of the geometric measures in the form of the L_p metric. The probabilistic measures used likelihood ratios that were derived from a Bayesian classifier that measured the relevancy of two images, one being the query image and one being a database image, so that image pairs which had a high likelihood value were classified as "relevant" and the ones which had a lower likelihood value were classified as "irrelevant". The first likelihood-based measure used multivariate Normal assumption and the second measure used independently fitted distributions for the feature differences. A classification-based approach with a minimum error decision rule was used to select the best performing p for the L_p metric. The values of p that resulted in the smallest classification errors and the largest average precisions were consistent and the classification scheme was an effective way of deciding which p to use. Experiments on a database of approximately 10,000 images showed that both likelihood-based measures performed significantly better than the commonly used L_p metrics in terms of average precision and recall. They were also more robust to normalization effects.

References

Abramowitz, M., Stegun, I.A. (Eds.), 1972. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards.

Aksoy, S., Haralick, R.M., 1998. Textural features for image database retrieval. In: Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, in conjunction with CVPR'98, Santa Barbara, CA, pp. 45–49.

Aksoy, S., Haralick, R.M., 2000a. Probabilistic vs. geometric similarity measures for image retrieval. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Hilton Head Island, SC, vol. 2, pp. 357–362.

Aksoy, S., Haralick, R.M., 2000b. Using texture in image similarity and retrieval. In: Pietikainen, M., Bunke, H. (Eds.), Texture Analysis in Machine Vision, volume 40 of Series in Machine Perception and Artificial Intelligence. World Scientific.

Belongie, S., Carson, C., Greenspan, H., Malik, J., 1988. Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In: Proc. IEEE International Conf. Computer Vision.

Berman, A.P., Shapiro, L.G., 1997. Efficient image retrieval with multiple distance measures. In: SPIE Storage and Retrieval of Image and Video Databases, pp. 12–21.

Bury, K.V., 1975. Statistical Models in Applied Science. Wiley, New York.

Casella, G., Berger, R.L., 1990. Statistical Inference. Duxbury Press, California.

Cheikh, F.A., Cramariuc, B., Reynaud, C., Qinghong, M., Adrian, B.D., Hnich, B., Gabbouj, M., Kerminen, P., Makinen, T., Jaakkola, H., 1999. Muvis: a system for content-based indexing and retrieval in large image databases. In: SPIE Storage and Retrieval of Image and Video Databases VII, San Jose, CA, pp. 98–106.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P., 1993. The QBIC project: querying images by content using color, texture and shape. In: SPIE Storage and Retrieval of Image and Video Databases. pp. 173–181.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall, NJ.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions, second ed. Wiley, New York, Vol. 2.

Li, C.S., Castelli, V., 1997. Deriving texture set for content-based retrieval of satellite image database. In: IEEE Internat. Conf. Image Processing. pp. 576–579.

Manjunath, B.S., Ma, W.Y., 1996. Texture features for browsing and retrieval of image data. IEEE Trans. Pattern Anal. Machine Intell. 18 (8), 837–842.

Nastar, C., Mitschke, M., Meilhac, C., 1998. Efficient query refinement for image retrieval. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Santa Barbara, CA, pp. 547–552.

Naylor, A.W., Sell, G.R., 1982. Linear Operator Theory in Engineering and Science. Springer, New York.

Papoulis, A., 1991. Probability, Random Variables, and Stochastic Processes, third edition. McGraw-Hill, New York.

Pentland, A., Picard, R.W., Sclaroff, S., 1994. Photobook: content-based manipulation of image databases. In: SPIE Storage and Retrieval of Image and Video Databases II. pp. 34–47.

Press, W.H., Flannary, B.P., Teukolsky, S.A., Vetterling, W.T., 1990. Numerical Recepies in C. Cambridge University Press, Cambridge.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.

Rui, Y., Huang, T.S., Chang, S.-F., 1999. Image retrieval: Current techniques, promising directions, and open issues. J. Visual Commun. Image Representation 10 (1), 39–62.

Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S., 1998. Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Trans. Circuits and Systems for Video Technology, Special Issue on Segmentation Description, and Retrieval of Video Content 8 (5), 644–655.

Sclaroff, S., Taycher, L., Cascia., M.L., 1997. Imagerover: a content-based image browser for the world wide web. In: Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries.

Smith, J.R., 1997. Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression. Ph.D. thesis, Columbia University.

Springer, M.D., 1979. The Algebra of Random Variables. Wiley, New York.