Modeling Image Similarity by Gaussian Mixture Models and the Signature Quadratic Form Distance

Christian Beecks, Anca Maria Ivanescu, Steffen Kirchhoff and Thomas Seidl
Data Management and Data Exploration Group, RWTH Aachen University, Germany
{beecks, ivanescu, kirchhoff, seidl}@cs.rwth-aachen.de

Abstract

Modeling image similarity for browsing and searching in voluminous image databases is a challenging task of nearly all content-based image retrieval systems. One promising way of defining image similarity consists in applying distance-based similarity measures on compact image representations. Beyond feature histograms and feature signatures, more general feature representations are mixture models of which the Gaussian mixture model is the most prominent one. This feature representation can be compared by employing approximations of the Kullback-Leibler Divergence. Although several of those approximations have been successfully applied to model image similarity, their applicability to mixture models based on high-dimensional feature descriptors is questionable. In this paper, we thus introduce the Signature Quadratic Form Distance to measure the distance between two Gaussian mixture models of high-dimensional feature descriptors. We show the analytical computation of the proposed Gaussian Quadratic Form Distance and evaluate its retrieval performance by making use of different benchmark image databases.

1. Introduction

Content-based multimedia retrieval [21] is a widespread and multifarious field of research which mainly focuses on the retrieval of complex multimedia data in voluminous databases. Including image, video, audio, or other non-text data, the challenging task of today's content-based multimedia retrieval systems is to support users in searching and exploring large multimedia databases. For this purpose, users frequently formalize their information needs by means of example objects or sketches reflecting the expected retrieval results, and the retrieval system is supposed to answer these user queries effectively and efficiently.

In general, given a user-specified query, content-based multimedia retrieval systems aim at providing users with the most similar multimedia objects contained in the database. This is done by first extracting the multimedia objects' inherent features and then computing a similarity value between the query and each database object in order to determine the most similar objects which have to be returned to the user. In particular, in the field of *content-based image retrieval* [7, 34], which focuses on modeling, management, and retrieval of images, similarity between two images is frequently determined by computing a distance between the corresponding *feature representations* which aggregate the images' inherent properties.

Prominent types of feature representation include *feature histograms* [30, 33, 35], *feature signatures* [3, 31], and *mixture models* [5, 38]. While feature histograms represent image properties in form of a vector, which is obtained by aggregating the features based on a *visual vocabulary*, i.e. a global partitioning, feature signatures aggregate the features of each image individually based on an adaptive partitioning. Feature signatures are thus free of a global visual vocabulary. Both feature histograms and feature signatures can be generalized by mixture models of which the *Gaussian mixture model* [6, 11, 27, 28] is frequently used in order to express image contents.

After having defined an appropriate feature representation, vectorial distances [15, 29] such as the *Minkowski* or *Quadratic Form Distance* [12], adaptive distances [3] such as the *Earth Mover's Distance* [31] or *Signature Quadratic Form Distance* [2, 4], and probabilistic distances [1, 9, 40] such as the *Kullback-Leibler Divergence* [20] are frequently encountered to determine image similarity between feature histograms, feature signatures, and components of mixture models, respectively.

While the Kullback-Leibler Divergence between two Gaussian densities can be computed by an analytic expression, that between two Gaussian mixture models can not be computed by a closed-form solution. Thus, besides the computationally intensive *Monte-Carlo* method, different approximations of the Kullback-Leibler Divergence have been proposed. Vasconcelos [37] presented a matching-based approach using the *Mahalanobis Distance* [23] which was improved by Goldberger et al. [10] using the Kullback-

2011 IEEE International Conference on Computer Vision
978-1-4577-1102-2/11/$26.00

Leibler Divergence to match two Gaussian densities. Furthermore, the same authors introduced a distance between Gaussian mixture models based on the *unscented transform* [18]. Finally, Hershey et al. [14] introduced a distance for Gaussian mixture models based on *variational methods* [17]. Although the aforementioned approximations of the Kullback-Leibler Divergence perform well, their ability to model image similarity based on high-dimensional feature descriptors [25, 36] is questionable.

In this paper, we present a novel method for modeling image similarity between Gaussian mixture models by making use of the Signature Quadratic Form Distance showing high retrieval performance among the major state-of-the-art adaptive similarity measures [3]. We introduce the Signature Quadratic Form Distance between Gaussian mixture models and show the analytical computation of this distance. In this way, we extend the applicability of the Signature Quadratic Form Distance to discrete and continuous feature distributions. By taking into account recent local feature descriptors of images, such as the *SIFT* [22] descriptor, we thoroughly evaluate the retrieval qualities regarding effectiveness and efficiency on different image databases and compare the effectiveness of our approach to the conventional *bag-of-visual-words* approach [33]. Our findings show that we significantly outperform the existing methods mentioned above in the context of content-based image retrieval with Gaussian mixture models.

We structure this paper as follows: in Section 2, we outline related work particularly regarding Gaussian mixture models and distances applicable to them. In Section 3, we introduce the Signature Quadratic Form Distance between Gaussian mixture models and show the analytical computation of this distance. We evaluate our approach on different benchmark image databases in Section 4 before we conclude our paper in Section 5.

2. Background and Related Work

Representing multimedia objects by means of features in a feature space is a challenging task of nearly all content-based analysis and retrieval approaches [7, 21, 34]. Whereas specific similarity matching tasks require the features to be accessible in an unaggregated way, content-based retrieval approaches generally require some degree of generalization in order to cope with different similarity notions. As a consequence, the features are first extracted and then aggregated by so-called *feature representations*.

While *bag-of-visual-words* approaches [33] aggregate the extracted features into a high-dimensional equi-length feature vector, i.e. a *feature histogram*, which can be compared by using vectorial distances [15, 29], alternative approaches tend to approximate the objects' contents through more flexible feature representations, so-called *feature signatures* [3, 31]. This type of feature representation adjusts to the contents of individual multimedia objects without using a global visual vocabulary and can be compared by making use of adaptive distance-based similarity measures [3], such as the Earth Mover's Distance [31] and the Signature Quadratic Form Distance [2, 4]. Although feature signatures show high applicability to any kind of local feature descriptors [25, 36], they typically aggregate the object's features according to a disjoint adaptive partitioning of these features which is frequently obtained by a clustering algorithm. This corresponds to the *hard assignment* [19] approach where each feature belongs to exactly one partition.

A more general approach, which unifies the concept of feature histograms and feature signatures is the *mixture model* [5, 38]. It models a feature distribution of an image by means of a mixture of densities. In this way, each feature histogram and feature signature can be expressed by a mixture of *Dirac delta functions*. The *Gaussian mixture model* [6, 11, 27, 28] is probably the most prominent one for image retrieval tasks and is defined over a multidimensional, vectorial, feature space $\mathbb{F} \subseteq \mathbb{R}^d$ as follows:

Definition 1 (Gaussian Mixture Model)
Given a feature space $\mathbb{F} \subseteq \mathbb{R}^d$. A Gaussian mixture model $g : \mathbb{F} \to \mathbb{R}$ is defined as:

$$g(x) = \sum_{i=1}^{n} w_i \cdot \mathcal{N}_{\mu_i,\Sigma_i}(x), \text{ where}$$

$$\mathcal{N}_{\mu_i,\Sigma_i}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} e^{-\frac{1}{2}(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)^T}$$

with prior probabilities $w_i \in \mathbb{R}$, means $\mu_i \in \mathbb{F}$, and covariance matrices $\Sigma_i \in \mathbb{R}^{d \times d}$.

The Gaussian mixture model defined above reflects the feature distribution of an image by a linear combination of Gaussian densities. In this way, the density of a single feature or the probability of regions in the feature space depends on all components $\mathcal{N}_{\mu_i,\Sigma_i}$ of the Gaussian mixture model. Thus, this model corresponds to the *soft assignment* [19] approach of modeling image contents.

Given an image $\mathcal{I}_a$, its Gaussian mixture model $g_a = \sum_{i=1}^{n} w_i^a \cdot \mathcal{N}_{\mu_i^a,\Sigma_i^a}$ can be computed as follows: first, the image features are extracted and represented in the corresponding feature space $\mathbb{F}$. For instance in case of extracting SIFT descriptors [22] the feature space comprises 128 dimensions reflecting gradient information around salient pixels, i.e. $\mathbb{F} \subseteq \mathbb{R}^{128}$. Second, the feature distribution is aggregated, for instance by applying the *expectation maximization* clustering algorithm [8]. As a result, each image is described by a Gaussian mixture model approximating the distribution of the corresponding image features. In the remainder of this paper, we will use the shorthand notation $\mathcal{N}_i^a$ instead of $\mathcal{N}_{\mu_i^a,\Sigma_i^a}$ where appropriate.

In order to measure the similarity between two images by computing a distance value between their feature rep-

resentations, the *Kullback-Leibler Divergence* [20] is frequently approximated when using Gaussian mixture models. In fact, the *Monte-Carlo* method approximates the Kullback-Leibler Divergence between two Gaussian mixture models g_a and g_b by taking a sufficiently large sampling $x_1, \ldots, x_n \in \mathbb{F}$ such that: $\mathrm{KL}(g_a||g_b) = \int g_a \log \frac{g_a}{g_b} \approx \frac{1}{n}\sum_{i=1}^{n} \log \frac{g_a(x_i)}{g_b(x_i)}$. Due to the high computational effort of this approximation, Goldberger et al. [10] proposed a matching-based approach between two Gaussian mixture models $g_a = \sum_{i=1}^{n} w_i^a \cdot \mathcal{N}_i^a$ and $g_b = \sum_{j=1}^{m} w_j^b \cdot \mathcal{N}_j^b$ which is defined as follows:

$$\mathrm{D}_{\mathrm{goldberger}}(g_a||g_b) = \sum_{i=1}^{n} w_i^a \left(\mathrm{KL}(\mathcal{N}_i^a||\mathcal{N}_{\pi(i)}^b) + \log \frac{w_i^a}{w_{\pi(i)}^b} \right),$$

where $\pi(i) = \arg\min_j \left(\mathrm{KL}(\mathcal{N}_i^a||\mathcal{N}_j^b - \log w_j^b \right)$ is the matching function between the components i and j of the Gaussian mixture models g_a and g_b, respectively. Based on this definition, Goldberger et al. improve the approach defined by Vasconcelos [37], who used the *Mahalanobis Distance* within the matching function π. Goldberger et al. also proposed another approach [10], which is based on the unscented transform:

$$\mathrm{D}_{\mathrm{unscented}}(g_a||g_b) = \frac{1}{2d} \left(\sum_{i=1}^{n} w_i^a \sum_{k=1}^{2d} \frac{\log g_a(x_{i,k})}{\log g_b(x_{i,k})} \right),$$

such that the sample points $x_{i,k} = \mu_i \pm \sqrt{d \cdot \lambda_{i,k}} \cdot e_{i,k}$ reflect the mean and variance of single components $\mathcal{N}_{\mu_i^a, \Sigma_i^a}$, while $\lambda_{i,k}$ and $e_{i,k}$ denote the i-th eigenvalue and eigenvector of Σ_i^a, respectively.

Whereas all of the aforementioned approaches are originally defined to model similarity in the context of content-based image retrieval, the next approach [14] was proposed and tested in the context of acoustic models used for speech recognition. It is defined between two Gaussian mixture models g_a and g_b as follows:

$$\mathrm{D}_{\mathrm{variational}}(g_a||g_b) = \sum_{i=1}^{n} w_i^a \log \left(\frac{\sum_{i'=1}^{n} w_{i'}^a e^{-\mathrm{KL}(\mathcal{N}_i^a||\mathcal{N}_{i'}^a)}}{\sum_{j=1}^{m} w_j^b e^{-\mathrm{KL}(\mathcal{N}_i^a||\mathcal{N}_j^b)}} \right).$$

Unlike the other approaches, $D_{variational}$ takes into account the correlation of components within the Gaussian mixture model g_a and that between the components of the Gaussian mixture models g_a and g_b.

To sum up, we briefly described approaches applicable to compare Gaussian mixture models. In particular, the approaches proposed by Goldberger et al. [10] and Vasconcelos [37] are designed to measure image similarity for image retrieval tasks. While these approaches follow the principle of approximating the well-natured Kullback-Leibler Divergence, we propose an orthogonal approach which makes use of the Signature Quadratic Form Distance in order to compare Gaussian mixture models.

3. Signature Quadratic Form Distance for Gaussian Mixture Models

In this section, we introduce the Signature Quadratic Form Distance between Gaussian mixture models. We first define the distance model between Gaussian mixture models and then propose a closed-form solution.

As introduced by Beecks et al. [2, 4], the Signature Quadratic Form Distance is an adaptive distance-based similarity measure defined for the comparison of flexible feature representations of different size and structure. Among the other state-of-the-art distances, it shows high retrieval performance in terms of effectiveness and efficiency [3]. It generalizes the well-known Quadratic Form Distance by adopting its *cross-dimension concept* in order to determine a distance value between two feature signatures which are sets of tuples comprising representatives and weights, i.e. each feature signature is defined as $\{\langle r_i, w_i \rangle \mid r_i \in \mathbb{F} \wedge w_i \in \mathbb{R}\}_{i=1}^{n}$ where $r_i \in \mathbb{F}$ denotes a representative in the underlying feature space $\mathbb{F}$ with corresponding weight $w_i \in \mathbb{R}$. The Signature Quadratic Form Distance adopts the cross-dimension concept of the Quadratic Form Distance by computing a similarity value between all representatives of the feature signatures. This is done by applying a so-called *similarity function* [4] which is defined over a feature space $\mathbb{F}$ as $f_s : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$. In particular, the Gaussian similarity function, defined as $f_s(x, y) = e^{-\alpha \cdot \mathrm{L}_2(x,y)^2}$ for a constant $\alpha \in \mathbb{R}^+$ and the Euclidean Distance $\mathrm{L}_2(x, y)$ between the two points $x, y \in \mathbb{F}$, has shown high retrieval quality [3]. As can be recognized, this similarity function behaves inversely proportional to the Euclidean Distance whose increase will result in a lower similarity value, and vice versa. After having defined an appropriate similarity function, the Signature Quadratic Form Distance will compare all representatives of both feature signatures with each other and multiply these similarities with the corresponding weights. In fact, the Signature Quadratic Form Distance between two finite feature signatures $S^a \subset \{\langle r^a, w^a \rangle | r^a \in \mathbb{F} \wedge w^a \in \mathbb{R}\}$ and $S^b \subset \{\langle r^b, w^b \rangle | r^b \in \mathbb{F} \wedge w^b \in \mathbb{R}\}$ is defined as follows: $\mathrm{SQFD}_{f_s}(S^a, S^b) = \sqrt{(w^a| - w^b) \cdot A_{f_s} \cdot (w^a| - w^b)^T}$ where $(w^a| - w^b)$ denotes the concatenation of weights from both feature signatures and A_{f_s} is the similarity matrix comprising similarity values between any pair of representatives of both feature signatures. Thus, the similarity matrix has to be dynamically computed for each distance computation between two feature signatures.

In order to compute the Signature Quadratic Form Distance between two Gaussian mixture models $g_a = \sum_{i=1}^{n} w_i^a \cdot \mathcal{N}_i^a$ and $g_b = \sum_{j=1}^{m} w_j^b \cdot \mathcal{N}_j^b$, we replace the similarity function between representatives of the feature signatures with the *expected similarity* between single components of the Gaussian mixture models. The expected similarity is defined as below.

Definition 2 (Expected Similarity)
Given two Gaussian densities $\mathcal{N}^a$ and $\mathcal{N}^b$ and a similarity function $f_s : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ over a feature space $\mathbb{F}$, the expected similarity $\mathrm{E}[f_s|\mathcal{N}^a, \mathcal{N}^b]$ of f_s with respect to $\mathcal{N}^a$ and $\mathcal{N}^b$ is defined as:

$$\mathrm{E}[f_s|\mathcal{N}^a, \mathcal{N}^b] = \int_{\mathbb{F}} \int_{\mathbb{F}} \mathcal{N}^a(x) \cdot \mathcal{N}^b(y) \cdot f_s(x, y) \, \mathrm{d}x\mathrm{d}y.$$

In other words, the similarity function f_s is interpreted as a random variable whose density is determined by the two Gaussian densities, and the expected similarity is thus the expected value of this random variable. In this way, the concept of expected similarity provides a very natural way of measuring similarity between two Gaussian densities. Based on the definition of expected similarity, we now define the *Gaussian Quadratic Form Distance* as the Signature Quadratic Form Distance between two Gaussian mixture models as below.

Definition 3 (Gaussian Quadratic Form Distance)
Given two Gaussian mixture models $g_a = \sum_{i=1}^{n} w_i^a \cdot \mathcal{N}_i^a$ and $g_b = \sum_{j=1}^{m} w_j^b \cdot \mathcal{N}_j^b$ and a similarity function $f_s : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ over a feature space $\mathbb{F}$, the Gaussian Quadratic Form Distance GQFD_{f_s} between g_a and g_b is defined as:

$$\mathrm{GQFD}_{f_s}(g_a, g_b) = \sqrt{(w^a| - w^b) \cdot A_{f_s} \cdot (w^a| - w^b)^T}$$

where $(w^a| - w^b) = (w_1^a, \ldots, w_n^a, -w_1^b, \ldots, -w_m^b)$ denotes the concatenation of prior probabilities from g_a and g_b, and $[a_{ij}] = A_{f_s} \in \mathbb{R}^{(n+m)\times(n+m)}$ is the similarity matrix whose entries are determined by computing the expected similarities between the corresponding components of the Gaussian mixture models as follows:

$$a_{ij} = \begin{cases} \mathrm{E}[f_s|\mathcal{N}_i^a, \mathcal{N}_j^a], & 1 \le i, j \le n, \\ \mathrm{E}[f_s|\mathcal{N}_i^a, \mathcal{N}_{j-n}^b], & 1 \le i \le n \ \wedge \ n+1 \le j \le n+m, \\ \mathrm{E}[f_s|\mathcal{N}_{i-n}^b, \mathcal{N}_j^a], & n+1 \le i \le n+m \ \wedge \ 1 \le j \le n, \\ \mathrm{E}[f_s|\mathcal{N}_{i-m+1}^b, \mathcal{N}_{j-m+1}^b], & n+1 \le i, j \le n+m. \end{cases}$$

By using a similarity function $f_s : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$, which is defined between any two points of the feature space $\mathbb{F}$, we can compute the Gaussian Quadratic Form Distance formalized in Definition 3 between two Gaussian mixture models. For this purpose, we have to compute the expected similarities between individual components of the Gaussian mixture models. The next lemma shows the closed-form expression of the expected similarity between two Gaussian densities with diagonal covariance matrices[1] when using the Gaussian similarity function.

[1] A diagonal covariance matrix is a matrix having values of zero beside the diagonal, i.e. $\forall i \neq j \ \sigma_{ii} \neq 0$ and $\sigma_{ij} = 0$.

Lemma 1 (Closed-Form Expression)
Given two Gaussian densities $\mathcal{N}_{\mu^a,\Sigma^a}$ and $\mathcal{N}_{\mu^b,\Sigma^b}$ with diagonal covariance matrices and the Gaussian similarity function $f_s(x, y) = e^{-\alpha \cdot \mathrm{L}_2(x,y)^2}$ over a feature space $\mathbb{F} \subseteq \mathbb{R}^d$, then the expected similarity formalized in Definition 2 can be simplified as follows:

$$\mathrm{E}[f_s|\mathcal{N}_{\mu^a,\Sigma^a}, \mathcal{N}_{\mu^b,\Sigma^b}] = \prod_{i=1}^{d} \frac{e^{-\frac{\alpha(\mu_i^a - \mu_i^b)^2}{1+2\alpha((\sigma_{ii}^a)^2 + (\sigma_{ii}^b)^2)}}}{\sqrt{1 + 2\alpha((\sigma_{ii}^a)^2 + (\sigma_{ii}^b)^2)}}$$

Proof 1
As we only consider multivariate Gaussian densities $\mathcal{N}_{\mu^a,\Sigma^a}$ and $\mathcal{N}_{\mu^b,\Sigma^b}$ with diagonal covariance matrices Σ^a and Σ^b, we can rewrite $\mathcal{N}_{\mu,\Sigma}$ as the product of its univariate Gaussian densities in each dimension, i.e. for $x \in \mathbb{F} \subseteq \mathbb{R}^d$ we have

$$\mathcal{N}_{\mu,\Sigma}(x) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i - \mu_i)^2}{\sigma_i^2}} = \prod_{i=1}^{d} \mathcal{N}_{\mu_i,\sigma_i}(x_i),$$

where $\mathcal{N}_{\mu_i,\sigma_i}(x_i)$ denotes the univariate Gaussian density in dimension i with mean μ_i and standard deviation $\sigma_i = \sigma_{ii}$. We further note that we can define the similarity function $f_s(x, y) = e^{-\alpha \cdot \mathrm{L}_2(x,y)^2}$ dimension-wise by

$$f_s(x, y) = \prod_{i=1}^{d} e^{-\alpha(x_i - y_i)^2} = \prod_{i=1}^{d} f_s^i(x_i, y_i),$$

where $x, y \in \mathbb{F} \subseteq \mathbb{R}^d$ are points in the feature space and $f_s^i : \mathbb{F}_i \times \mathbb{F}_i \to \mathbb{R}$ denotes the Gaussian similarity function applied to a single dimension i of the feature space. Then,

$$\begin{aligned} & \mathrm{E}[f_s|\mathcal{N}_{\mu^a,\Sigma^a}, \mathcal{N}_{\mu^b,\Sigma^b}] \\ = & \int_{\mathbb{F}} \int_{\mathbb{F}} \mathcal{N}_{\mu^a,\Sigma^a}(x) \cdot \mathcal{N}_{\mu^b,\Sigma^b}(y) \cdot f_s(x, y) \, \mathrm{d}x\mathrm{d}y \\ = & \prod_{i=1}^{d} \int_{\mathbb{F}_i} \int_{\mathbb{F}_i} \mathcal{N}_{\mu_i^a,\sigma_i^a}(x_i) \cdot \mathcal{N}_{\mu_i^b,\sigma_i^b}(y_i) \cdot f_s^i(x_i, y_i) \mathrm{d}x_i\mathrm{d}y_i \\ = & \prod_{i=1}^{d} \int_{\mathbb{F}_i} \mathcal{N}_{\mu_i^a,\sigma_i^a}(x_i) \cdot \left(\int_{\mathbb{F}_i} \mathcal{N}_{\mu_i^b,\sigma_i^b}(y_i) \cdot f_s^i(x_i, y_i) \mathrm{d}y_i \right) \mathrm{d}x_i. \end{aligned}$$

We will first solve the inner integral by showing that for every dimension i it holds:

$$\int_{\mathbb{F}_i} \mathcal{N}_{\mu_i^b,\sigma_i^b}(y_i) \cdot f_s^i(x_i, y_i) \mathrm{d}y_i = \frac{1}{\sqrt{1 + 2\alpha(\sigma_i^b)^2}} e^{\frac{-\alpha(x_i - \mu_i^b)^2}{1+2\alpha(\sigma_i^b)^2}}.$$

We have

$$\begin{aligned} & \int_{\mathbb{F}_i} \mathcal{N}_{\mu_i^b,\sigma_i^b}(y_i) \cdot f_s^i(x_i, y_i) \mathrm{d}y_i \\ = & \int_{\mathbb{F}_i} \frac{1}{\sqrt{2\pi}\sigma_i^b} e^{-\frac{1}{2}\frac{(y_i - \mu_i^b)^2}{(\sigma_i^b)^2}} \cdot e^{-\alpha(x_i - y_i)^2} \mathrm{d}y_i \end{aligned}$$

$$= \int_{\mathbb{F}_i} \frac{1}{\sqrt{2\pi}\sigma_i^b} e^{-\left(\frac{1}{2(\sigma_i^b)^2}+\alpha\right)y_i^2+\left(\frac{\mu_i^b}{(\sigma_i^b)^2}+2\alpha x_i\right)y_i+\left(\frac{-(\mu_i^b)^2}{2(\sigma_i^b)^2}-\alpha x_i^2\right)} \mathrm{d}y_i.$$

By substituting $k = \frac{1}{\sqrt{2\pi}\sigma_i^b}$, $f = \frac{1}{2(\sigma_i^b)^2}+\alpha$, $g = \frac{\mu_i^b}{(\sigma_i^b)^2}+2\alpha x_i$, *and* $h = \frac{-(\mu_i^b)^2}{2(\sigma_i^b)^2}-\alpha x_i^2$, *we can solve the Gaussian integral by*

$$\begin{aligned}\int_{\mathbb{F}_i} k e^{-fy_i^2+gy_i+h}\mathrm{d}y_i &= \int_{\mathbb{F}_i} k e^{-f(y_i-\frac{g}{2f})^2+\frac{g^2}{4f}+h}\mathrm{d}y_i \\ &= k\cdot\sqrt{\frac{\pi}{f}}e^{\frac{g^2}{4f}+h} \\ &= \frac{1}{\sqrt{1+2\alpha(\sigma_i^b)^2}} e^{\frac{-\alpha(x_i-\mu_i^b)^2}{1+2\alpha(\sigma_i^b)^2}}.\end{aligned}$$

This integral converges as f *is strictly positive (*α *and* σ_i^b *are positive). Analogously, we can solve the outer integral:*

$$\begin{aligned}&\prod_{i=1}^{d}\int_{\mathbb{F}_i}\mathcal{N}_{\mu_i^a,\sigma_i^a}(x_i)\cdot\left(\int_{\mathbb{F}_i}\mathcal{N}_{\mu_i^b,\sigma_i^b}(y_i)\cdot f_s^i(x_i,y_i)\mathrm{d}y_i\right)\mathrm{d}x_i \\ =&\prod_{i=1}^{d}\int_{\mathbb{F}_i}\frac{1}{\sqrt{2\pi}\sigma_i^a}e^{-\frac{1}{2}\frac{(x_i-\mu_i^a)^2}{(\sigma_i^a)^2}}\cdot\left(\frac{1}{\sqrt{1+2\alpha(\sigma_i^b)^2}}e^{\frac{-\alpha(x_i-\mu_i^b)^2}{1+2\alpha(\sigma_i^b)^2}}\right)\mathrm{d}x_i \\ =&\prod_{i=1}^{d}\int_{\mathbb{F}_i}k' e^{-f'x_i^2+g'x_i+h'}\mathrm{d}x_i,\end{aligned}$$

with $k' = \frac{1}{\sqrt{2\pi(\sigma_i^a)^2(1+2\alpha(\sigma_i^b)^2)}}$, $f' = \frac{1}{2(\sigma_i^a)^2}+\frac{\alpha}{1+2\alpha(\sigma_i^b)^2}$, $g' = \frac{\mu_i^a}{(\sigma_i^a)^2}+\frac{2\alpha\mu_i^b}{1+2\alpha(\sigma_i^b)^2}$, *and* $h' = \frac{-(\mu_i^a)^2}{2(\sigma_i^a)^2}-\frac{\alpha(\mu_i^b)^2}{1+2\alpha(\sigma_i^b)^2}$. *This finally yields*

$$\mathrm{E}[f_s|\mathcal{N}_{\mu^a,\Sigma^a},\mathcal{N}_{\mu^b,\Sigma^b}] = \prod_{i=1}^{d}\frac{e^{-\frac{\alpha(\mu_i^a-\mu_i^b)^2}{1+2\alpha((\sigma_i^a)^2+(\sigma_i^b)^2)}}}{\sqrt{1+2\alpha((\sigma_i^a)^2+(\sigma_i^b)^2)}}.$$

□

As proved in Lemma 1, the expected similarity between two Gaussian densities with diagonal covariance matrices can be computed dimension-wise by computing the applied Gaussian similarity function for each dimension individually. In this way, we can compute the Gaussian Quadratic Form Distance between two Gaussian mixture models comprising diagonal covariance matrices by using the closed-form expression of the expected similarity. In the next section, we will examine the effectiveness and efficiency of the proposed approach in the context of content-based image retrieval using high-dimensional feature descriptors.

4. Experimental Evaluation

In this section, we compare the retrieval performance of the proposed Gaussian Quadratic Form Distance (GQFD_{f_s}) using the Gaussian similarity function $f_s(x,y) = e^{-\frac{\mathrm{L}_2(x,y)^2}{1.0E5}}$ with that of the other approaches: $\mathrm{D}_{\mathrm{variational}}$ [14], $\mathrm{D}_{\mathrm{unscented}}$ [10], and $\mathrm{D}_{\mathrm{goldberger}}$ [10]. In addition, we compare the results to the conventional bag-of-visual-words approach. To this end, we make use of the following benchmark image databases: the *Corel Wang* database [39], the *UCID* database [32], and the *Coil100* database [26]. All image databases provide a ground truth that is used to determine the effectiveness of the aforementioned approaches.

The *Corel Wang* database comprises 1,000 images from 10 different classes such as *beaches*, *busses*, *flowers*, etc. Although images belonging to the same class correspond to the topic of that particular class, this image database shows the highest *heterogeneity* with respect to visual similarity within one class and is thus considered as a difficult benchmark database. The *UCID* database contains over 1,300 uncompressed color images designed for benchmarking image retrieval approaches. The provided ground truth of this database takes into account the strong visual correlation between query images and those database images belonging to the same class the individual query image belongs to. Thus, it is a *homogeneous* database regarding visual similarity of images from the same class, and the number of different images in each class is below 18. The *Coil100* database comprises 7,200 images classified into 100 classes. Each class contains images depicting the same object which is photographed from 72 different perspectives. Here, the ground truth is strongly motivated by the object which is depicted in the image.

Based on the aforementioned image databases, we generated a Gaussian mixture model comprising 10 components for each single image. This was done by extracting the following local feature descriptors: *RGBhistogram* (45d), which reflects color information of a single image pixel, and the more complex feature descriptors *SIFT* (128d), *CSIFT* (384d), *RGSIFT* (384d), *RGBSIFT* (384d), and *OpponentSIFT* (384d), which reflect aggregated gradient information around an image pixel. These feature descriptors were extracted by dense sampling, and the *expectation maximization* clustering algorithm was applied to cluster the extracted feature descriptors of each image individually. For this purpose, we made use of the color descriptor software provided by van de Sande et al. [36] to extract local feature descriptors and the WEKA software [13] to generate the Gaussian mixture models.

In order to evaluate the retrieval effectiveness, we measured *mean average precision values* [24] of complete database rankings with respect to 100 distinct random queries regarding each combination of image database and feature descriptor. The results are depicted in Tables 1, 2, and 3 for the *Corel Wang*, *UCID*, and *Coil100* database.

As can be seen in the tables, the Gaussian Quadratic Form Distance (GQFD_{f_s}) achieves the highest mean av-

Table 1. Mean average precision values for the *Corel Wang* database.

	GQFD_{f_s}	$\mathrm{D}_{\mathrm{variational}}$[14]	$\mathrm{D}_{\mathrm{unscented}}$[10]	$\mathrm{D}_{\mathrm{goldberger}}$[10]
OpponentSIFT	0.398	0.277	0.196	0.179
RGBSIFT	0.407	0.248	0.196	0.173
RGSIFT	0.399	0.264	0.196	0.177
CSIFT	0.399	0.269	0.196	0.172
SIFT	0.457	0.264	0.278	0.177
RGBhistogram	0.368	0.260	0.259	0.259

Table 2. Mean average precision values for the *UCID* database.

	GQFD_{f_s}	$\mathrm{D}_{\mathrm{variational}}$[14]	$\mathrm{D}_{\mathrm{unscented}}$[10]	$\mathrm{D}_{\mathrm{goldberger}}$[10]
OpponentSIFT	0.555	0.285	0.008	0.105
RGBSIFT	0.534	0.310	0.008	0.085
RGSIFT	0.573	0.246	0.008	0.124
CSIFT	0.576	0.249	0.008	0.097
SIFT	0.529	0.352	0.471	0.159
RGBhistogram	0.503	0.432	0.486	0.427

Table 3. Mean average precision values for the *Coil100* database.

	GQFD_{f_s}	$\mathrm{D}_{\mathrm{variational}}$[14]	$\mathrm{D}_{\mathrm{unscented}}$[10]	$\mathrm{D}_{\mathrm{goldberger}}$[10]
OpponentSIFT	0.401	0.134	0.021	0.092
RGBSIFT	0.365	0.149	0.021	0.109
RGSIFT	0.427	0.156	0.021	0.100
CSIFT	0.434	0.133	0.021	0.092
SIFT	0.450	0.124	0.036	0.099
RGBhistogram	0.600	0.235	0.309	0.231

erage precision values for each combination of image database and feature descriptor. This can be explained by the well-natured behavior of the Gaussian Quadratic Form Distance which takes into account all possible expected similarities of both Gaussian mixture models. The expected similarities of the Gaussian Quadratic Form Distance are not only computed between components of two different Gaussian mixture models but also between components within each Gaussian mixture model. Therefore, the computation of the Gaussian Quadratic Form Distance depends on the complete structure of both mixture models, which indeed is crucial for high-dimensional data. A similar behavior can be recognized for $\mathrm{D}_{\mathrm{variational}}$. This approach also takes into account the Kullback-Leibler Divergence between all components of the query Gaussian mixture model. Besides the Gaussian Quadratic Form Distance, it achieves the highest mean average precision values for all image databases regarding complex and high-dimensional feature descriptors. While $\mathrm{D}_{\mathrm{variational}}$ shows higher retrieval quality in terms of mean average precision values than $\mathrm{D}_{\mathrm{goldberger}}$ and $\mathrm{D}_{\mathrm{unscented}}$, it is generally outperformed by the latter when using *SIFT* and *RGBhistogram* feature descriptors. In these cases, $\mathrm{D}_{\mathrm{unscented}}$ reaches acceptable high mean average precision values in the *UCID* database compared to the Gaussian Quadratic Form Distance. We state that this approach making use of the unscented transform only achieves high retrieval quality when the feature descriptor's dimensionality is low. Using the 384-dimensional feature descriptors *CSIFT*, *RGSIFT*, *RGBSIFT*, and *OpponentSIFT*, this approach frequently fails because of the computation of the logarithmic Gaussian mixture density $\log\left(\prod_{i=1}^{2*384} g(x_{i,k})\right)$ which results in a value of minus infinity when some point $x_{i,k}$ has a probability density value of zero. Due to algorithmic reasons, this problem was usually encountered when using high-dimensional feature descriptors. Therefore, the computed $\mathrm{D}_{\mathrm{unscented}}$ value is meaningless and the mean average precision values stay constant in each database using these high-dimensional feature descriptors. The other approaches were not affected by this algorithmic problem.

Regarding the retrieval performance, we state that the Gaussian Quadratic Form Distance (GQFD_{f_s}) shows the highest retrieval quality in terms of mean average precision values. Using feature descriptors which are not high-dimensional, $\mathrm{D}_{\mathrm{unscented}}$ achieves comparable retrieval quality. For high-dimensional descriptors, $\mathrm{D}_{\mathrm{variational}}$ shows the best results.

Let us now compare the results of our approach to that of the bag-of-visual-words approach. Fur this purpose, we used the *SIFT* descriptors of the *Flickr60k* database [16] in order to generate visual vocabularies comprising 100 to

Table 4. Mean average precision values for the bag-of-visual-words approach.

	100	200	500	1k	2k	5k	10k	20k	50k	100k	200k	GQFD_{f_s}
Corel Wang	0.450	0.389	0.399	0.421	0.427	0.425	0.437	0.461	0.463	0.451	0.428	0.457
UCID	0.520	0.504	0.514	0.538	0.544	0.547	0.549	0.554	0.547	0.541	0.537	0.529
Coil100	0.396	0.400	0.428	0.442	0.434	0.442	0.435	0.428	0.413	0.417	0.400	0.450

Table 5. Computation time values for 1,000 distance computations in milliseconds.

	GQFD_{f_s}	$\mathrm{D}_{\mathrm{variational}}$[14]	$\mathrm{D}_{\mathrm{unscented}}$[10]	$\mathrm{D}_{\mathrm{goldberger}}$[10]
**SIFT* (384d)	1,307	53	421,774	34
SIFT (128d)	458	50	50,018	31
RGBhistogram (45d)	181	47	7,241	25

200,000 visual words. The results regarding mean average precision values are reported in Table 4, where we used the Euclidean Distance to compare the feature histograms. As can be seen in the table, the number of visual words needed to reach the retrieval performance of the GQFD_{f_s} depends on the individual image database. While in the *Corel Wang* database the bag-of-visual-words approach using 20k and 50k visual words shows slightly higher mean average precision values than our approach, in the *Coil100* database the retrieval performance of the bag-of-visual-words approach always stays below that of the GQFD_{f_s}.

Let us finally investigate the computation time needed to compare two Gaussian mixture models. In Table 5, we report these computation time values in milliseconds needed for 1,000 distance computations. As the computation time values are approximately the same for all high-dimensional feature descriptors, we summarize these feature descriptors by **SIFT*. As can be seen in the table, the computation time values correlate with the size of the feature descriptors. In particular, both $\mathrm{D}_{\mathrm{variational}}$ and $\mathrm{D}_{\mathrm{goldberger}}$ have the lowest computation times which stay below 53 milliseconds for any kind of feature descriptor. While these approaches are extremely fast, $\mathrm{D}_{\mathrm{unscented}}$ shows the highest computation time which is at least above 7 seconds for the feature descriptor *RGBhistogram*. This is due to the unscented transform which generates 2d many points that have to be generated for each computation individually. For this reason we conclude that $\mathrm{D}_{\mathrm{unscented}}$ is inappropriate for the comparison of Gaussian mixture models of high-dimensional feature spaces and thus feature descriptors. Our proposed approach, the Gaussian Quadratic Form Distance (GQFD_{f_s}), needs approximately 0.2 and 1.3 seconds to perform 1,000 distance computations on the low-dimensional feature descriptor *RGBhistogram* and the high-dimensional feature descriptor **SIFT*, respectively, which is acceptable compared to the retrieval quality regarding effectiveness.

To sum up, we have shown that the introduced Gaussian Quadratic Form Distance combines high retrieval effectiveness with acceptable retrieval efficiency even on high-dimensional feature descriptors. Unlike existing approaches, the proposed distance is able to compute meaningful distance values between high-dimensional Gaussian mixture models for the purpose of content-based image retrieval. Therefore, the Gaussian Quadratic Form Distance is an appropriate distance-based similarity measure to model content-based similarity between images.

5. Conclusions and Future Work

We have introduced the Signature Quadratic Form Distance for the comparison of Gaussian mixture models in the context of content-based image retrieval. For this purpose, we have provided a closed-form expression in order to compute this distance analytically. By modeling image contents through high-dimensional local feature descriptors, we have shown that the proposed *Gaussian Quadratic Form Distance* is able to outperform state-of-the-art approaches in terms of retrieval effectiveness and that it shows acceptable retrieval efficiency.

As future work, we plan to approximate and index the Gaussian Quadratic Form Distance in order to improve the efficiency of content-based query processing. Furthermore, we plan to examine the applicability of the other state-of-the-art adaptive similarity measures in order to model image similarity based on Gaussian mixture models.

Acknowledgments The authors gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB-686.

References

[1] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Process.*, 18:349–369, 1989. 1

[2] C. Beecks, M. S. Uysal, and T. Seidl. Signature quadratic form distances for content-based similarity. In *Proc. ACM Multimedia*, pages 697–700, 2009. 1, 2, 3

[3] C. Beecks, M. S. Uysal, and T. Seidl. A comparative study of similarity measures for content-based multimedia retrieval. In *Proc. IEEE ICME*, pages 1552–1557, 2010. 1, 2, 3

[4] C. Beecks, M. S. Uysal, and T. Seidl. Signature quadratic form distance. In *Proc. ACM CIVR*, pages 438–445, 2010. 1, 2, 3

[5] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE TPAMI*, 29:394–410, 2007. 1, 2

[6] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE TPAMI*, 24:1026–1038, 2002. 1, 2

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008. 1, 2

[8] A. Dempster, N. Laird, D. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. 2

[9] P. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982. 1

[10] J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Proc. IEEE ICCV*, pages 487–493, 2003. 1, 3, 5, 6, 7

[11] H. Greenspan, J. Goldberger, and L. Ridel. A continuous probabilistic framework for image matching. *Comput. Vis. Image Underst.*, 84:384–406, 2001. 1, 2

[12] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE TPAMI*, 17(7):729–736, 1995. 1

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, 2009. 5

[14] J. Hershey and P. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *Proc. IEEE ICASSP*, volume 4, pages 317–320, 2007. 2, 3, 5, 6, 7

[15] R. Hu, S. Rüger, D. Song, H. Liu, and Z. Huang. Dissimilarity measures for content-based image retrieval. In *Proc. IEEE ICME*, pages 1365 –1368, 2008. 1, 2

[16] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, pages 304–317, 2008. 6

[17] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999. 2

[18] S. Julier and J. Uhlmann. A general method for approximating nonlinear transformations of probability distributions. *Dept. of Engineering Science, University of Oxford, Tech. Rep*, 1996. 2

[19] M. J. Kearns, Y. Mansour, and A. Y. Ng. *An information-theoretic analysis of hard and soft assignment methods for clustering*, pages 495–520. MIT Press, Cambridge, MA, USA, 1999. 2

[20] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):pp. 79–86, 1951. 1, 3

[21] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):1–19, 2006. 1, 2

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, 2004. 2

[23] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, 1936. 1

[24] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. 5

[25] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE TPAMI*, 27(10):1615–1630, 2005. 2

[26] S. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-100). Technical report, Department of Computer Science, Columbia University, 1996. 5

[27] H. Permuter, J. Francos, and I. Jermyn. Gaussian mixture models of texture and colour for image database retrieval. In *Proc. IEEE ICASSP*, volume 3, pages 569 – 572, 2003. 1, 2

[28] H. Permuter, J. Francos, and I. Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recogn.*, 39:695–706, 2006. 1, 2

[29] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Comput. Vis. Image Underst.*, 84:25–43, 2001. 1, 2

[30] Y. Rubner and C. Tomasi. *Perceptual Metrics for Image Database Navigation*. Kluwer Academic Publishers, Norwell, MA, USA, 2001. 1

[31] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vision*, 40(2):99–121, 2000. 1, 2

[32] G. Schaefer and M. Stich. UCID: an uncompressed color image database. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5307, pages 472–480, 2003. 5

[33] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. IEEE ICCV*, pages 1470–1477, 2003. 1, 2

[34] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE TPAMI*, 22:1349–1380, 2000. 1, 2

[35] M. J. Swain and D. H. Ballard. Color indexing. *Int. J. Comput. Vision*, 7:11–32, 1991. 1

[36] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE TPAMI*, 32(9):1582–1596, 2010. 2, 5

[37] N. Vasconcelos. On the Complexity of Probabilistic Image Retrieval. In *Proc. IEEE ICCV*, volume 2, pages 400–407, 2001. 1, 3

[38] N. Vasconcelos and A. Lippmann. Feature representations for image retrieval: Beyond the color histogram. In *Proc. IEEE ICME*, pages 899–902, 2000. 1, 2

[39] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE TPAMI*, 23(9):947–963, 2001. 5

[40] S. K. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. *IEEE TPAMI*, 28:917–929, 2006. 1