Signature Matching Distance for Content-based Image Retrieval

Christian Beecks
RWTH Aachen University
beecks@cs.rwth-aachen.de

Steffen Kirchhoff
Harvard University
kirchhoff@fas.harvard.edu

Thomas Seidl
RWTH Aachen University
seidl@cs.rwth-aachen.de

ABSTRACT

We propose a simple yet effective approach to content-based image retrieval: the *signature matching distance*. While recent approaches to content-based image retrieval utilize the bag-of-visual-words model, where image descriptors are matched through a common visual vocabulary, signature-based approaches use a distance between signatures, i.e. between image-specific bags of locally aggregated descriptors, in order to quantify image dissimilarity. In this paper, we focus on the signature-based approach to content-based image retrieval and propose a novel distance function, the signature matching distance. This distance matches coincident visual properties of images based on their signatures. In particular, by investigating different descriptor matching strategies and their suitability to match signatures, we show that our approach is able to outperform other signature-based approaches to content-based image retrieval. Moreover, in combination with a simple color and texture-based image descriptor, our approach is able to compete with the majority of bag-of-visual-words approaches.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Search process; I.4.10 [**Image Processing and Computer Vision**]: Image representation

General Terms

Algorithms, Experimentation, Performance

Keywords

content-based image retrieval, distance function, feature signature, signature matching distance

1. INTRODUCTION

Content-based image retrieval [5, 20] denotes the process of retrieving images from an image database that share similar contents with respect to a given query image or query sketch. The probably most frequently employed approach to perform this task is the *bag-of-visual-words* (BoVW) model [19], which has been further investigated in recent years in order to maximize the retrieval performance. Jégou et al. [8] introduced *Hamming embedding*, which has been enhanced by Jain et al. [7] through an asymmetric version. Other approaches are the *vectors of locally aggregated descriptors* [11] and the *compressed Fisher vectors* [16], which are particularly designed for large-scale image retrieval. In general, these approaches outperform the conventional BoVW model by improving the visual vocabulary through a more descriptive quantization of the underlying feature space. This comes along with a computational extensive preprocessing phase. Besides the extraction of local feature descriptors, such as SIFT [12], a large and descriptive visual vocabulary is learned in order to model image similarity more precisely.

An alternative approach to content-based image retrieval is the *signature-based* model. Instead of describing all images by means of the same visual vocabulary, signature-based approaches represent each image individually through an image-specific visual vocabulary, which can be adapted to specific image properties dynamically, even at runtime. The resulting image representations, namely the signatures, can then be compared by distance-based dissimilarity measures [1, 17] such as the *earth mover's distance* [18] or the *signature quadratic form distance* [2].

If one is willing to model image similarity in a content-based way, either by the BoVW model or the signature-based model, one of the major challenges lies in the definition of a suitable similarity model that concentrates on the characteristic features of the images and facilitates an efficient image retrieval process. According to the latest findings [7, 11], comparing visual features of images by matching their corresponding image descriptors along a visual vocabulary has shown to achieve the highest retrieval performance in terms of efficiency and accuracy.

In this paper, we take advantage of this matching-based similarity definition and incorporate it into the signature-based model. We introduce the *signature matching distance* (SMD) as a novel approach to content-based image retrieval. The SMD adaptively defines the dissimilarity between two images based on a matching between their signatures and thus without the necessity of a common visual vocabulary. As a result, the proposed SMD is able to outperform other signature-based approaches to content-based image retrieval in terms of efficiency and accuracy. Moreover, by using a simplistic color and texture-based image descriptor, our performance evaluation on the Holidays [8] and UKBench [14] benchmark image databases reveals that the SMD is able to compete with the majority of BoVW approaches.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMR'13, April 16–20, 2013, Dallas, Texas, USA.
Copyright 2013 ACM 978-1-4503-2033-7/13/04 ...$15.00.

2. APPROACHES TO CONTENT-BASED IMAGE RETRIEVAL

In order to carry out the process of content-based image retrieval, images are modeled mathematically. This is done by first describing characteristic image properties through local feature descriptors, such as SIFT [12], and then by quantizing these descriptors in order to derive a compact and indexable image representation. While *bag-of-visual-words approaches* (BoVW) use the same quantization for all images of a database, *signature-based approaches* use a specific quantization for each image individually.

2.1 BoVW Approaches

Sivic and Zisserman [19] came up with the idea of modeling image similarity with the bag-of-visual-words model. The idea of this model is to define a similarity value between images by means of their visual word occurrences with respect to a *visual vocabulary*. Thus, an image $\mathcal{I}$ with descriptors $f(\mathcal{I}) = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ is described by assigning each descriptor x_i to its nearest visual word $q(x_i)$. The occurrences of the visual words are then counted and additionally weighted by a *tf-idf* weighting scheme such that an image $\mathcal{I}$ is finally described by the BoVW-vector $v_\mathcal{I} = (\text{tf-idf}(s_1), \ldots, \text{tf-idf}(s_k))^T \in \mathbb{R}^k$, where $s_j = |q^{-1}(j) \cap f(\mathcal{I})|$ denotes the number of image descriptors that are assigned to visual word j for $1 \leq j \leq k$. Typically, the similarity between two images $\mathcal{I}$ and $\mathcal{J}$ is then modeled by the normalized dot product $\text{BoVW}(v_\mathcal{I}, v_\mathcal{J}) = \frac{v_\mathcal{I}^T \cdot v_\mathcal{J}}{|v_\mathcal{I}| \cdot |v_\mathcal{J}|}$.

Jégou et al. [8] enhanced the conventional BoVW model by the *Hamming embedding*. The basic idea of this approach is to increase the descriptiveness of the visual vocabulary by additionally taking into account the descriptor distributions along the visual words. For this purpose, each descriptor x_i of an image $\mathcal{I}$ is not only assigned to a visual word $q(x_i)$ but also to a binary signature $b(x_i) \in \{0,1\}^{d_b}$ that approximates the location of the descriptor within the Voronoi cell of $q(x_i)$. Thus, an image $\mathcal{I}$ with descriptors $f(\mathcal{I}) = \{x_i\}_{i=1}^n$ is quantized by the set of tuples $x = \{(q(x_i), b(x_i))\}_{i=1}^n$. The similarity between two images $\mathcal{I}$ and $\mathcal{J}$ with descriptor quantizations $x = \{(q(x_i), b(x_i))\}_{i=1}^n$ and $y = \{(q(y_j), b(y_j))\}_{j=1}^m$ is then defined by $\text{HE}(x,y) = \frac{1}{\sqrt{m}} \sum_{i=1}^n \sum_{j=1}^m f_{\text{HE}}(x_i, y_j)$ with the Hamming embedding matching function $f_{\text{HE}}(x_i, y_j) = \text{tf-idf}(q(x_i))^2$ if and only if $q(x_i) = q(y_j) \wedge \text{H}(b(x_i), b(y_j)) \leq h_t$ and zero otherwise. Here, H denotes the Hamming distance and $h_t \in \mathbb{R}$ a fixed Hamming threshold with $0 \leq h_t \leq d_b$.

Jain et al. [7] introduced the idea of an *asymmetric Hamming embedding* by replacing the Hamming distance $\text{H}(b(x_i), b(y_j))$ between the binary signatures $b(x_i)$ and $b(y_j)$ of two descriptors x_i and y_j that are assigned to the same visual word $q(x_i) = q(y_j)$ with a sum of asymmetric distances that reflect the proximity of the descriptors more precisely.

Perronnin et al. [16] exploited the *Fisher kernel framework* [6] in order to describe each image by its gradients of the log-likelihood function given a common generative model. In this way, the generative model is used as a visual vocabulary. By using a *Gaussian mixture model* $p(x) = \sum_{i=1}^k \pi_i \cdot \mathcal{N}_{\mu_i, \Sigma_i}(x)$ with diagonal covariance matrices Σ_i, each image $\mathcal{I}$ with descriptors $f(\mathcal{I}) \subset \mathbb{R}^d$ is then represented by its *Fisher vector* $\mathcal{G}_\mathcal{I} = (G_1, \ldots, G_k)^T \in \mathbb{R}^{k \cdot d}$ with components $G_i = \frac{1}{\sqrt{\pi_i}} \sum_{x \in f(\mathcal{I})} \text{P}(i|x) \cdot \Sigma^{-1} \cdot (x - \mu_i) \in \mathbb{R}^d$, where $\text{P}(i|x) = \frac{\pi_i \cdot \mathcal{N}_{\mu_i, \Sigma_i}(x)}{\sum_{j=1}^k \pi_j \cdot \mathcal{N}_{\mu_j, \Sigma_j}(x)}$ denotes the soft assignment of descriptor x to visual word i, which is modeled by the Gaussian density $\mathcal{N}_{\mu_i, \Sigma_i}$ for $1 \leq i \leq k$. The similarity between two images $\mathcal{I}$ and $\mathcal{J}$ is then defined by the dot product $\text{FV}(\mathcal{G}^\mathcal{I}, \mathcal{G}^\mathcal{J}) = \mathcal{G}_\mathcal{I}^T \cdot \mathcal{G}_\mathcal{J}$. The authors further propose to binarize/normalize these vectors by using *power normalization*, *locality sensitive hashing* [4], or *spectral hashing* [23].

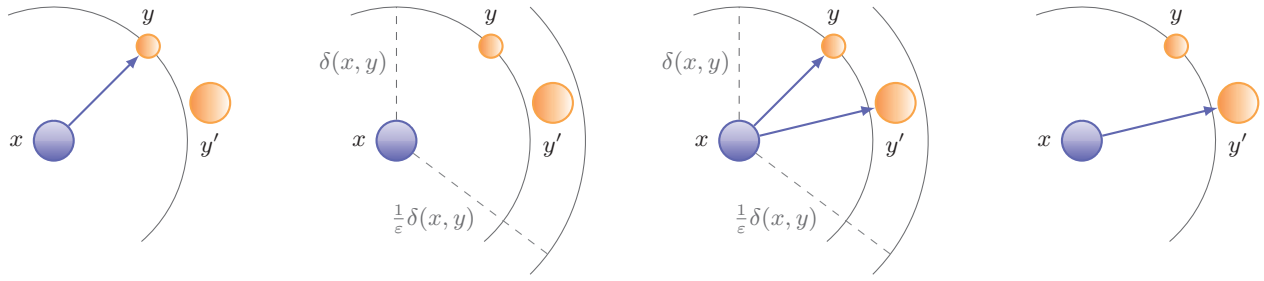
Jégou et al. [11] introduced the *vector of locally aggregated descriptors* (VLAD) as a simplified non-probabilistic version of the Fisher vector. By assuming equal probabilities $\pi_i = \frac{1}{k}$ and *isotropic* covariance matrices Σ_i, the Fisher vector of each image $\mathcal{I}$ with descriptors $f(\mathcal{I}) \subset \mathbb{R}^d$ becomes the VLAD-descriptor $\mathcal{V}_\mathcal{I} = (V_1, \ldots, V_k)^T \in \mathbb{R}^{k \cdot d}$ with components $V_i = \sum_{x \in f(\mathcal{I}) \wedge q(x) = i} (x - \mu_i) \in \mathbb{R}^d$, where each mean $\mu_i \in \mathbb{R}^d$ is the visual word representing $q(x) = i$ for $1 \leq i \leq k$. Given two images $\mathcal{I}$ and $\mathcal{J}$, their distance is defined by $\text{VLAD}(\mathcal{V}_\mathcal{I}, \mathcal{V}_\mathcal{J}) = \text{L}_2(\mathcal{V}_\mathcal{I}, \mathcal{V}_\mathcal{J})$. Further, normalizations of the VLAD-descriptors have been proposed [11].

2.2 Signature-based Approaches

Signature-based approaches quantize the descriptors of each image individually by means of an image-specific visual vocabulary, namely the *signature* [18]. Given a descriptor space $\mathbb{R}^d$, a signature X is defined by a finite set of representatives $\text{R}_X \subset \mathbb{R}^d$, where each representative is additionally assigned to a non-negative weight by a weighting function $w_X : \text{R}_X \to \mathbb{R}^{\geq 0}$. Thus, a signature X can be defined mathematically as the graph of its weighting function w_X, i.e. $X = \{(x, w_X(x)) | x \in \text{R}_X\}$. Given an image $\mathcal{I}$, its signature can be computed by clustering the set of extracted descriptors $f(\mathcal{I})$ and defining the representatives of the signature by the cluster centroids. The weighting function can be defined by the corresponding cluster sizes. In this way, the representatives and weights correspond to the visual words and their frequencies of an image-specific visual vocabulary. Since these visual words differ, signature-based approaches apply a so-called *ground distance* $\delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ between all representatives in order to define a *signature distance*.

The *earth mover's distance* (EMD) [18] is a *transformation-based* approach measuring the cost of transforming one signature into another. The earth mover's distance EMD_δ between two signatures X and Y is defined as a minimum cost flow over all possible flows $[f_{xy}] = F \in \mathbb{R}^{|\text{R}_X| \times |\text{R}_Y|}$ between the elements $x, y \in \text{R}_X \cup \text{R}_Y$, i.e. $\text{EMD}_\delta(X,Y) = \min_F \left\{ \frac{\sum_{x \in \text{R}_X} \sum_{y \in \text{R}_Y} f_{xy} \cdot \delta(x,y)}{\min\{\sum_{x \in \text{R}_X} w_X(x), \sum_{y \in \text{R}_Y} w_Y(y)\}} \right\}$, subject to the constraints $\forall x, y : f_{xy} \geq 0$, $\forall x \in \text{R}_X : \sum_{y \in \text{R}_Y} f_{xy} \leq w_X(x)$, $\forall y \in \text{R}_Y : \sum_{x \in \text{R}_X} f_{xy} \leq w_Y(y)$, and $\sum_{x \in \text{R}_X} \sum_{y \in \text{R}_Y} f_{xy} = \min\{\sum_{x \in \text{R}_X} w_X(x), \sum_{y \in \text{R}_Y} w_Y(y)\}$.

The *perceptually modified Hausdorff distance* (PMHD) [15] is a *matching-based* approach. A *matching* between two signatures X, Y can be defined as $\text{m}_{X \to Y} = \{(x, \pi_{X \to Y}(x)) | \forall x \in \text{R}_X\}$, where the *matching function* $\pi_{X \to Y} : \text{R}_X \to \text{R}_Y$ maps each representative $x \in \text{R}_X$ to one representative $y \in \text{R}_Y$. Additionally, a cost function $\text{c} : 2^{\text{R}_X \times \text{R}_Y} \to \mathbb{R}$ defines the cost of a matching. The perceptually modified Hausdorff distance is then defined as $\text{PMHD}_\delta(X,Y) = \max\{\text{c}(\text{m}_{X \to Y}), \text{c}(\text{m}_{Y \to X})\}$. The matching $\text{m}_{X \to Y}$ is defined by the graph of the matching function $\pi_{X \to Y}(x) = \text{argmin}_{y \in \text{R}_Y} \left\{ \frac{\delta(x,y)}{\min\{w_X(x), w_Y(y)\}} \right\}$, and the cost c of the matching is defined as $\text{c}(\text{m}_{X \to Y}) = \sum_{(x,y) \in \text{m}_{X \to Y}} \frac{w_X(x)}{\sum_{(x,y) \in \text{m}_{X \to Y}} w_X(x)} \cdot \frac{\delta(x,y)}{\min\{w_X(x), w_Y(y)\}}$. The matching $\text{m}_{Y \to X}$ is defined analogously.



(a) $\mathrm{m}^{\mathrm{NN}}_{X\to Y} = \{(x,y)\}$ (b) $\mathrm{m}^{\delta\mathrm{r}}_{X\to Y} = \emptyset$ (c) $\mathrm{m}^{\mathrm{i}\delta\mathrm{r}}_{X\to Y} = \{(x,y),(x,y')\}$ (d) $\mathrm{m}^{\delta/w^*}_{X\to Y} = \{(x,y')\}$

Figure 1: Illustration of the different matching strategies between two signatures $X = \{(x, w_X(x))\}$ and $Y = \{(y, w_Y(y)), (y', w_Y(y'))\}$: (a) nearest neighbor matching, (b) distance ratio matching, (c) inverse distance ratio matching, and (d) distance weight ratio matching.

The *signature quadratic form distance* (SQFD) [2] is a *correlation-based* approach that uses a *similarity function* $\mathrm{s} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to quantify similarities among representatives of the signatures. Based on a similarity function s, we define the *weighted similarity correlation* between two signatures X and Y as $X \cdot_{\mathrm{s}} Y = \sum_{x\in \mathrm{R}_X} \sum_{y\in \mathrm{R}_Y} w_X(x) \cdot w_Y(y) \cdot \mathrm{s}(x,y)$. The signature quadratic form distance $\mathrm{SQFD}_{\mathrm{s}}$ is then defined as follows: $\mathrm{SQFD}_{\mathrm{s}}(X,Y) = \sqrt{X \cdot_{\mathrm{s}} X - X \cdot_{\mathrm{s}} Y - Y \cdot_{\mathrm{s}} X + Y \cdot_{\mathrm{s}} Y}$.

To sum up, BoVW approaches utilize a common visual vocabulary in order to match image descriptors to predetermined visual words. In contrast, signature-based approaches utilize image-specific visual vocabularies in order to adapt the distance computation to individual descriptor distributions. Depending on the intended notion of similarity, different types of distances can thus be used. For instance for the purpose of content-based image retrieval where image similarity is frequently assessed by the amount of properties the images share, it is meaningful to attribute the distance computation to the most similar parts of the signatures. This is done by a *matching*, as investigated in the next section.

3. SIGNATURE MATCHING STRATEGIES

In this section, we present different matching strategies and discuss their suitability for modeling similarity between signatures. Given two signatures X and Y with their representatives R_X and R_Y, a *matching* $\mathrm{m}_{X\to Y}$ between X and Y is defined as a subset of the Cartesian product of their representatives, i.e. $\mathrm{m}_{X\to Y} \subseteq \mathrm{R}_X \times \mathrm{R}_Y$. This definition includes the assignment of multiple representatives from R_X to R_Y and vice versa. The most intuitive way to match representatives between signatures is by means of the concept of the nearest neighbor, which is defined for an element $x \in \mathbb{R}^d$, a ground distance δ, and a set $\mathrm{R} \subset \mathbb{R}^d$ as $\mathrm{NN}_{\delta,\mathrm{R}}(x) = \{y \mid y = \mathrm{argmin}_{y'\in \mathrm{R}} \delta(x,y')\}$. This leads to the following definition of the *nearest neighbor matching* [13].

Definition 1. *Nearest Neighbor Matching*
Given two signatures X, Y and a ground distance δ, the nearest neighbor matching $\mathrm{m}^{\mathrm{NN}}_{X\to Y} \subseteq \mathrm{R}_X \times \mathrm{R}_Y$ *from X to Y is defined as:*

$$\mathrm{m}^{\mathrm{NN}}_{X\to Y} = \{(x,y) \mid y \in \mathrm{NN}_{\delta,\mathrm{R}_Y}(x)\}.$$

The nearest neighbor matching $\mathrm{m}^{\mathrm{NN}}_{X\to Y}$ satisfies both *left totality* and *right uniqueness*, i.e. $\forall x \in \mathrm{R}_X \exists y \in \mathrm{R}_Y : (x,y) \in \mathrm{m}^{\mathrm{NN}}_{X\to Y}$ and $\forall x \in \mathrm{R}_X, \forall y,z \in \mathrm{R}_Y : (x,y) \in \mathrm{m}^{\mathrm{NN}}_{X\to Y} \wedge (x,z) \in \mathrm{m}^{\mathrm{NN}}_{X\to Y} \Rightarrow y = z$. Each representative $x \in \mathrm{R}_X$ is matched to the representative $y \in \mathrm{R}_Y$ that minimizes $\delta(x,y)$. Thus, the nearest neighbor matching $\mathrm{m}^{\mathrm{NN}}_{X\to Y}$ is of size $|\mathrm{m}^{\mathrm{NN}}_{X\to Y}| = |\mathrm{R}_X|$. Figure 1(a) provides an example of $\mathrm{m}^{\mathrm{NN}}_{X\to Y}$ where $x \in \mathrm{R}_X$ is matched to $y \in \mathrm{R}_Y$. As can be seen in this example, the distance $\delta(x,y)$ and $\delta(x,y')$ differs only marginally. Thus, the nearest neighbor matching becomes ambiguous, since both representatives y and y' serve as good matching candidates.

A well-known strategy to overcome the issue of ambiguity of the nearest neighbor matching is given by the *distance ratio matching* [13]. Intuitively, it is defined by matching only those parts of the signatures that are unique with respect to the ratio of the nearest and second nearest neighbor.

Definition 2. *Distance Ratio Matching*
Given two signatures X, Y and a ground distance δ, the distance ratio matching $\mathrm{m}^{\delta\mathrm{r}}_{X\to Y} \subseteq \mathrm{R}_X \times \mathrm{R}_Y$ *from X to Y is defined by a parameter $\varepsilon \in [0,1]$ as:*

$$\mathrm{m}^{\delta\mathrm{r}}_{X\to Y} = \{(x,y) \mid y \in \mathrm{NN}_{\delta,\mathrm{R}_Y}(x) \wedge y' \in \mathrm{NN}_{\delta,\mathrm{R}_Y\setminus\{y\}}(x) \wedge \frac{\delta(x,y)}{\delta(x,y')} < \varepsilon\}.$$

The distance ratio matching $\mathrm{m}^{\delta\mathrm{r}}_{X\to Y}$ does not satisfy left totality but it satisfies right uniqueness, i.e. every $x \in \mathrm{R}_X$ is matched to at most one $y \in \mathrm{R}_Y$. Thus, the size of this matching is $|\mathrm{m}^{\delta\mathrm{r}}_{X\to Y}| \leq |\mathrm{R}_X|$. In the extreme case, the matching could even be empty, i.e. $\mathrm{m}^{\delta\mathrm{r}}_{X\to Y} = \emptyset$, as illustrated in Figure 1(b). In fact, the matching $\mathrm{m}^{\delta\mathrm{r}}_{X\to Y}$ epitomizes a defensive matching strategy. It completely rejects those pairs that result in an ambiguous matching. Since we empirically found that this behavior jeopardizes the retrieval performance, see Section 5, we propose a more offensive strategy that works the other way around. Instead of excluding those pairs (x,y) and (x,y') from $\mathrm{m}^{\mathrm{NN}}_{X\to Y}$ that cause ambiguity, we propose to include them, as formalized in the definition below.

Definition 3. *Inverse Distance Ratio Matching*
Given two signatures X, Y and a ground distance δ, the inverse distance ratio matching $\mathrm{m}^{\mathrm{i}\delta\mathrm{r}}_{X\to Y} \subseteq \mathrm{R}_X \times \mathrm{R}_Y$ *from X to Y is defined by a parameter $\varepsilon \in [0,1]$ as:*

$$\mathrm{m}^{\mathrm{i}\delta\mathrm{r}}_{X\to Y} = \mathrm{m}^{\mathrm{NN}}_{X\to Y} \cup \{(x,y') \mid y \in \mathrm{NN}_{\delta,\mathrm{R}_Y}(x) \wedge y' \in \mathrm{NN}_{\delta,\mathrm{R}_Y\setminus\{y\}}(x) \wedge \frac{\delta(x,y)}{\delta(x,y')} > \varepsilon\}.$$

In contrast to the distance ratio matching, the inverse variant $\mathrm{m}^{\mathrm{i}\delta\mathrm{r}}_{X\to Y}$ satisfies left totality but not right uniqueness. As can be seen in Figure 1(c), each $x \in \mathrm{R}_X$ is assigned to

at most two representatives from signature Y. This leads to a matching size of $|\mathrm{m}^{\mathrm{i\delta r}}_{X\to Y}| \leq 2 \cdot |\mathrm{R}_X|$. In general, it can be shown that the inverse distance ratio matching is a generalization of the nearest neighbor matching, while the distance ratio matching is a specialization, i.e. it holds that $\mathrm{m}^{\delta\mathrm{r}}_{X\to Y} \subseteq \mathrm{m}^{\mathrm{NN}}_{X\to Y} \subseteq \mathrm{m}^{\mathrm{i\delta r}}_{X\to Y}$ with equality for $\varepsilon = 1$.

The aforementioned matchings consider only the spatial distance δ between two representatives $x \in \mathrm{R}_X$ and $y \in \mathrm{R}_Y$ in a descriptor space $\mathbb{R}^d$, not their weights $w_X(x)$ and $w_Y(y)$. These weights may nonetheless significantly contribute to the similarity definition. Thus, we propose the *distance weight ratio matching*, which is defined in a self-adjusting manner for two representatives $x \in \mathrm{R}_X$ and $y \in \mathrm{R}_Y$ by means of their *distance weight ratio* $\delta\!/\!w^*(x,y) = \delta(x,y)/\min\{w_X(x), w_Y(y)\}$. By utilizing this ratio $\delta\!/\!w^*$, we define the distance weight ratio matching free of any predefined threshold as it is used for the calculation of the PMHD.

Definition 4. *Distance Weight Ratio Matching*
Given two signatures X, Y and a ground distance δ, the distance weight ratio matching $\mathrm{m}^{\delta/w^*}_{X\to Y} \subseteq \mathrm{R}_X \times \mathrm{R}_Y$ *from X to Y is defined as:*

$$\mathrm{m}^{\delta/w^*}_{X\to Y} = \{(x,y) \,|\, y = \mathrm{argmin}_{y'\in \mathrm{R}_Y} \delta\!/\!w^*(x,y')\}.$$

The distance weight ratio matching $\mathrm{m}^{\delta/w^*}_{X\to Y}$ satisfies both left totality and right uniqueness. Thus, this matching is of size $|\mathrm{m}^{\delta/w^*}_{X\to Y}| = |\mathrm{R}_X|$. In addition to only minimizing the distance between representatives, this matching also considers the weights of the representatives. By dividing the ground distance δ between two representatives x and y by their minimal weight, this matching penalizes those representatives y that have a smaller weight than representative x. This is illustrated in Figure 1(d). Although representative y is located slightly closer to x than representative y', x is not matched to y since the fact that $w_Y(y) < w_X(x)$ increases the distance weight ratio such that $\delta\!/\!w^*(x,y') < \delta\!/\!w^*(x,y)$. As a consequence, the distance weight ratio matching suppresses the contribution of noisy representatives of the signatures.

We have presented four different matching strategies that allow to match coincident visual properties of signatures. Based on these matchings, we introduce the *signature matching distance* in the following section.

4. SIGNATURE MATCHING DISTANCE

In this section, we introduce our novel signature-based approach to content-based image retrieval, the *signature matching distance* (SMD). This will be complemented by an analysis of its properties and suitable cost functions that provide a theoretical means of assessing the quality of a matching.

4.1 Definition

The fundamental idea of the SMD is to model the distance between two signatures by means of the cost of the symmetric difference of the matching elements of the signatures. In general, the *symmetric difference* $A \Delta B$ of two sets A and B is the set of elements which are contained in either A or B but not in their intersection $A \cap B$, i.e. $A \Delta B = A \cup B \setminus A \cap B$. By adapting this concept to matchings between signatures X and Y, the set A becomes the matching $\mathrm{m}_{X\to Y}$ from signature X to Y, and the set B becomes the matching $\mathrm{m}_{Y\to X}$ from signature Y to X. The symmetric difference $\mathrm{m}_{X\to Y} \Delta \mathrm{m}_{Y\to X} = \{(x,y) | (x,y) \in \mathrm{m}_{X\to Y} \oplus (y,x) \in \mathrm{m}_{Y\to X}\}$

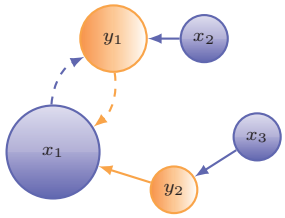


Figure 2: Matching-based principle of the SMD between two signatures $X = \{(x_i, w_X(x_i))\}_{i=1}^3$ and $Y = \{(y_i, w_Y(y_i))\}_{i=1}^2$. While the symmetric difference $\mathrm{m}_{X\to Y} \Delta \mathrm{m}_{Y\to X}$ completely neglects the matching between x_1 and y_1, the SMD includes this bidirectional matching dependent on parameter λ.

can then be evaluated by a cost function in order to model a distance between the corresponding signatures X and Y.

An example of the symmetric difference $\mathrm{m}_{X\to Y} \Delta \mathrm{m}_{Y\to X}$ between two signatures $X = \{(x_i, w_X(x_i))\}_{i=1}^3$ and $Y = \{(y_i, w_Y(y_i))\}_{i=1}^2$ is depicted in Figure 2, where the representatives of X and Y are shown by blue and orange circles, and the corresponding weights are indicated by the respective diameters. In this example, the distance weight ratio matching defines the matchings $\mathrm{m}^{\delta/w^*}_{X\to Y} = \{(x_1,y_1),(x_2,y_1),(x_3,y_2)\}$ and $\mathrm{m}^{\delta/w^*}_{Y\to X} = \{(y_1,x_1),(y_2,x_1)\}$, which are depicted by blue and orange arrows between the corresponding representatives of the signatures. As can be seen in the figure, the symmetric difference $\mathrm{m}_{X\to Y} \Delta \mathrm{m}_{Y\to X} = \{(x_2,y_1),(x_3,y_2),(x_1,y_2)\}$ completely neglects bidirectional matches that are depicted by the dashed arrows, i.e. it neglects those pairs of representatives $x \in \mathrm{R}_X$ and $y \in \mathrm{R}_Y$ for which holds that $(x,y) \in \mathrm{m}_{X\to Y} \wedge (y,x) \in \mathrm{m}_{Y\to X}$.

On the one hand, excluding these bidirectional matches corresponds to the idea of measuring dissimilarity by those elements of the signatures that are less similar, on the other hand the exclusion of bidirectional matches reduces the discriminability of similar signatures whose matchings mainly comprise bidirectional matches. In order to balance this trade-off, we generalize the symmetric difference and define the SMD with an additional real-valued parameter $\lambda \in [0,1]$ that models the *exclusion of bidirectional matchings* from the distance computation.

Definition 5. *Signature Matching Distance*
Given two signatures X and Y, a matching strategy m*, and a cost function* c*, the* signature matching distance SMD *between X and Y with respect to parameter $\lambda \in [0,1]$ is defined as follows:*

$$\mathrm{SMD}_\lambda(X,Y) = \mathrm{c}(\mathrm{m}_{X\to Y}) + \mathrm{c}(\mathrm{m}_{Y\to X}) - 2\lambda \cdot \mathrm{c}(\mathrm{m}_{X\leftrightarrow Y}).$$

The computation of the SMD between two signatures X and Y is carried out by adding the costs $\mathrm{c}(\mathrm{m}_{X\to Y})$ and $\mathrm{c}(\mathrm{m}_{Y\to X})$ of matching representatives from X to Y and from Y to X, and subtracting the cost $\mathrm{c}(\mathrm{m}_{X\leftrightarrow Y})$ of the corresponding bidirectional matching $\mathrm{m}_{X\leftrightarrow Y} = \{(x,y) | (x,y) \in \mathrm{m}_{X\to Y} \wedge (y,x) \in \mathrm{m}_{Y\to X}\}$. The cost $\mathrm{c}(\mathrm{m}_{X\leftrightarrow Y})$ are multiplied by parameter $\lambda \in [0,1]$ and doubled, since bidirectional matches occur in both matchings $\mathrm{m}_{X\to Y}$ and $\mathrm{m}_{Y\to X}$. As mentioned above, the parameter λ models the exclusion of bidirectional matchings. In particular, a value of $\lambda = 0$ includes the cost of bidirectional matchings in the distance computation, while a value of $\lambda = 1$ excludes the cost of bidirectional matchings in the distance computation. In case $\lambda = 1$ the SMD between two signatures X and Y becomes

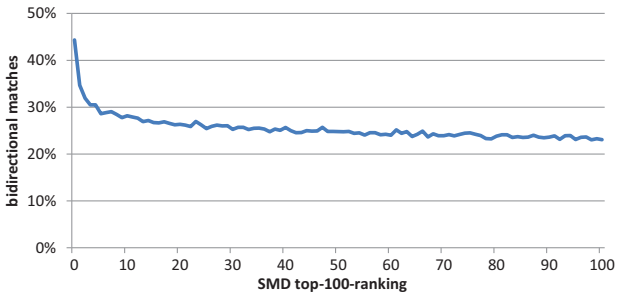


Figure 3: Percentage of bidirectional matches for the distance weight ratio matching m^{δ/w^*} on the Holidays database for the 100 most similar images.

the cost of the symmetric difference of the corresponding matchings, i.e. $\mathrm{SMD}_{\lambda=1}(X,Y) = \mathrm{c}(\mathrm{m}_{X\to Y} \,\Delta\, \mathrm{m}_{Y\to X})$.

But which particular role play bidirectional matches for the purpose of content-based image retrieval? We briefly discuss their impact on the dissimilarity definition before we continue with describing one prominent way of assessing the cost of a matching by means of a ground distance.

4.2 Bidirectional Matches

Intuitively, bidirectional matches describe a more stable relationship between two representatives of the corresponding signatures than unidirectional matches. Consequently, the number of bidirectional matches between two signatures should increase with increasing similarity of the corresponding images. We could observe this intuition empirically on the Holidays image database by measuring the percentage of bidirectional matches on basis of the distance weight ratio matching m^{δ/w^*}. As can be seen in Figure 3, the average percentage of bidirectional matches for the 100 most similar images regarding multiple queries decreases with increasing rank of the images. While the matching between the query and its nearest neighbor on average comprises 44% bidirectional matches, this number diminishes to 23% for the images at ranking position 100. Further, the least similar images contain only 10% bidirectional matches.

As a consequence, adjusting the parameter λ to a value of smaller than 1 in order to include the cost of bidirectional matches seems to be appropriate for most similarity definitions. We will investigate the choice and stability of the parameter λ in Section 5.

4.3 Cost Functions

The calculation of the SMD between two signatures follows a two-step approach. First, a matching between the representatives of the signatures is determined. Second, this matching is evaluated according to a cost function. A naive way of assessing the cost of a matching consists in counting the number of matches. This would, however, only roughly reflect the similarity relationship between two signatures, because the cardinality of a matching does not take into account the coincident characteristics of the matching elements. An alternative that considers the differences of the matching elements are ground distance-based cost functions.

Definition 6. *Ground Distance-based Cost Functions Given a matching $\mathrm{m}_{X\to Y}$ between two signatures X and Y and a ground distance δ, we define the following* ground distance-based cost functions $\mathrm{c} : 2^{\mathrm{R}_X \times \mathrm{R}_Y} \to \mathbb{R}^{\geq 0}$*:*

$$\mathrm{c}_\delta(\mathrm{m}_{X\to Y}) = \sum_{(x,y)\in \mathrm{m}_{X\to Y}} w_X(x)\cdot w_Y(y)\cdot \delta(x,y),$$

$$\mathrm{c}_{\delta/w^*}(\mathrm{m}_{X\to Y}) = \sum_{(x,y)\in \mathrm{m}_{X\to Y}} w_X(x)\cdot w_Y(y)\cdot \delta/w^*(x,y).$$

The idea of both ground distance-based cost functions is to evaluate a matching $\mathrm{m}_{X\to Y}$ by its weighted ground distances δ or its weighted distance weight ratios δ/w^* between the matching representatives x and y of the signatures X and Y. The more similar the matching representatives, the smaller the values of δ and δ/w^*, thus the lower the cost of the respective matching. Intuitively, provided that the weights of the signatures are normalized, i.e. $\sum_{x\in \mathrm{R}_X} w_X(x) = \sum_{y\in \mathrm{R}_Y} w_Y(y) = 1$, these cost functions are equivalent to the expected ground distance and to the expected distance weight ratio constrained to the matching $\mathrm{m}_{X\to Y}$, i.e. $\mathrm{c}_\delta(\mathrm{m}_{X\to Y}) = \mathrm{E}[\delta|\mathrm{m}_{X\to Y}]$ and $\mathrm{c}_{\delta/w^*}(\mathrm{m}_{X\to Y}) = \mathrm{E}[\delta/w^*|\mathrm{m}_{X\to Y}]$, respectively. The normalization of the signature weights is necessary in order to not favor smaller matchings over larger matchings.

To sum up, the SMD is a matching-based distance function designed for signatures. Provided that the computation time complexity of the underlying matching $\mathrm{m}_{X\to Y}$ lies in $\mathcal{O}(|\mathrm{R}_X|\cdot|\mathrm{R}_Y|)$ and that the cost function can be computed in linear time complexity with respect to the matching size, i.e. for a given matching $\mathrm{m}_{X\to Y}$ between two signatures X and Y its time complexity lies in $\mathcal{O}(|\mathrm{m}_{X\to Y}|)$, the SMD has a quadratic computation time complexity of $\mathcal{O}(|\mathrm{R}_X|\cdot|\mathrm{R}_Y|)$.

5. PERFORMANCE EVALUATION

In this section, we study the performance of the SMD and compare it to that of state-of-the-art content-based image retrieval approaches. For this purpose, we used the Holidays [8] and UKBench [14] image databases, both providing a solid ground truth for benchmarking image retrieval approaches. The Holidays database comprises 1,491 holiday photos corresponding to a large variety of scene types. It was designed to test the robustness, for instance, to rotation, viewpoint, and illumination changes and provides 500 selected queries. The UKBench database consists of 10,200 images showing 2,550 different objects or scenes that are photographed from four different viewpoints. The first image of each object or scene serves as query object.

Based on these databases, we generated signatures by extracting local feature descriptors and by clustering them with the k-means algorithm. We extracted three different descriptors: a low-dimensional descriptor denoted by PCT [1], which is based on position, color, and texture, as well as the SIFT [12] and CSIFT [3] descriptors. The PCT descriptor describes the relative spatial information of a pixel, its CIELAB color value, and its first and second Tamura texture features [21], which are coarseness and contrast. We utilized a random sampling of 40,000 pixels to extract the PCT descriptors and the Harris-Laplace detector to extract the SIFT and CSIFT descriptors. The color descriptor software provided by van de Sande et al. [22] was used to extract the SIFT and CSIFT descriptors. After having extracted the local feature descriptors, we applied the k-means clustering algorithm to generate multiple signatures per image by varying the signature size between 10 and 100.

Table 1: Retrieval performance of the SMD on the Holidays and UKBench databases by using different combinations of matching strategies and descriptors. The ground distance is set to $\delta = L_1$, the cost function is set to c_{δ/w^*}, and bidirectional matches are not excluded, i.e. $\lambda = 0$. The parameter $\varepsilon \in [0, 1]$ defines the threshold of the matchings $m^{\delta r}$ and $m^{i\delta r}$.

		Holidays			UKBench			
		MAP	ε	size	MAP	score	ε	size
m^{NN}	PCT	0.810	–	100	0.845	3.20	–	60
	SIFT	0.653	–	40	0.463	1.75	–	20
	CSIFT	0.735	–	20	0.531	2.00	–	20
$m^{\delta r}$	PCT	0.810	1.0	100	0.845	3.20	1.0	60
	SIFT	0.653	1.0	40	0.463	1.75	1.0	20
	CSIFT	0.735	1.0	20	0.531	2.00	1.0	20
$m^{i\delta r}$	PCT	**0.822**	0.8	70	**0.860**	3.27	0.7	60
	SIFT	0.663	0.6	40	0.517	1.98	0.8	90
	CSIFT	0.755	0.8	30	0.591	2.23	0.8	20
m^{δ/w^*}	PCT	0.819	–	90	0.855	3.24	–	100
	SIFT	0.656	–	40	0.463	1.74	–	20
	CSIFT	0.725	–	20	0.529	2.00	–	20

5.1 Performance Analysis of the SMD

The retrieval performance of the SMD is summarized in Table 1, where we report the *mean average precision* (MAP) values for the Holidays and UKBench databases and the *score* [14] for the latter. The score denotes the average number of relevant images ranked within the four nearest neighbors of a query. The retrieval performance is based on all queries defined by the ground truth of the corresponding databases. We additionally include the optimal values of parameter $\varepsilon \in [0, 1]$, which defines the threshold of the matchings $m^{\delta r}$ and $m^{i\delta r}$. The reported values were obtained by using the distance weight ratio cost function c_{δ/w^*} with the Manhattan ground distance L_1 since we empirically found that this combination outperforms the other variants.

As can be seen in Table 1, both databases reveal the same tendencies. First, the SMD on PCT-based signatures outperforms that on SIFT-based and CSIFT-based signatures by reaching the highest MAP values of 0.822 and 0.860 on the Holidays and UKBench databases with signatures of sizes 70 and 60, respectively. Second, the distance ratio matching $m^{\delta r}$ provides acceptable retrieval performance only when setting its inherent threshold ε to a value of 1, thus turning this matching into the nearest neighbor matching, i.e. $m^{NN} = m^{\delta r}$. Third, the retrieval performance of the aforementioned matchings and of the distance weight ratio matching m^{δ/w^*} is slightly below that of our proposed inverse distance ratio matching $m^{i\delta r}$, which mitigates the issue of matching ambiguity.

We continue with investigating the stability of the SMD with respect to the parameters $\varepsilon \in [0, 1]$ and $\lambda \in [0, 1]$ that model the matching threshold and the influence of bidirectional matches. Figures 4 and 5 depict the retrieval performance for the Holidays and UKBench databases by utilizing the matching $m^{i\delta r}$ in combination with PCT-based signatures, since this combination has shown the highest retrieval performance. As can be seen in Figure 4(a) and Figure 5(a), the threshold ε of the matching $m^{i\delta r}$ is less sensitive to changes the larger the signature size, and vice versa. By setting the matching threshold ε to an approximate value of 0.7, which corresponds to the empirical observation of Mikolajczyk and Schmid [13], and using PCT-based signatures of size greater than 30, the retrieval performance of the SMD stays above a MAP value of 0.8 for both databases. In general, the retrieval performance of the SMD is further improved by excluding the costs of bidirectional matches, as shown in Figure 4(b) and Figure 5(b). By increasing the parameter λ to the values of 0.85 and 0.90, the SMD achieves the MAP values of 0.834 and 0.876 on the Holidays and UKBench databases, respectively.

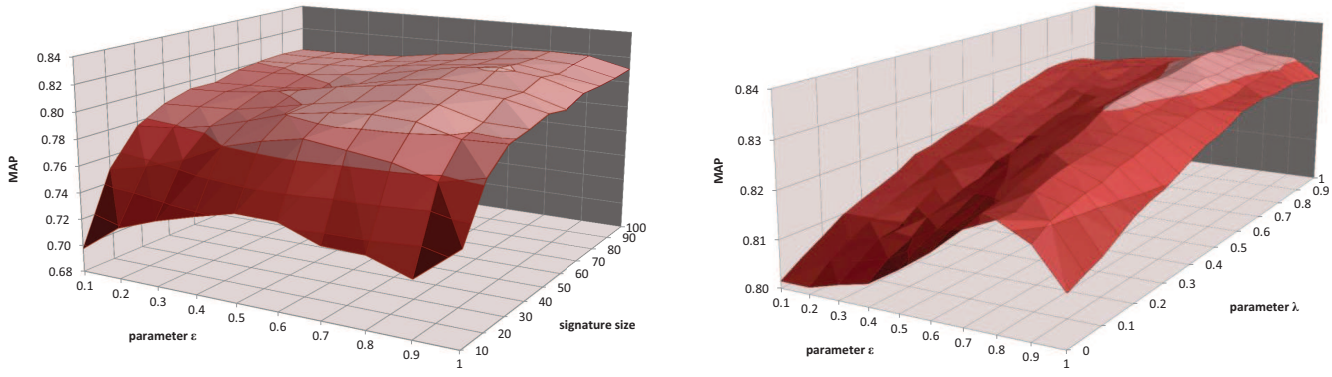
Table 2: Retrieval performance of the signature-based approaches on the Holidays and UKBench databases.

		Holidays		UKBench		
		MAP	size	MAP	score	size
EMD	PCT	0.720	90	0.741	2.78	50
	SIFT	0.678	70	0.536	2.05	90
	CSIFT	0.749	40	0.605	2.31	30
PMHD	PCT	0.804	80	**0.866**	3.30	90
	SIFT	0.673	70	0.531	2.03	90
	CSIFT	0.755	40	0.594	2.27	30
SQFD	PCT	0.761	40	0.766	2.86	60
	SIFT	0.690	80	0.585	2.23	100
	CSIFT	0.756	20	0.494	2.25	20
SMD	PCT	**0.810**	100	0.845	3.20	60
	SIFT	0.653	40	0.463	1.75	20
	CSIFT	0.735	20	0.531	2.00	20

5.2 Comparison to Signature-based Approaches

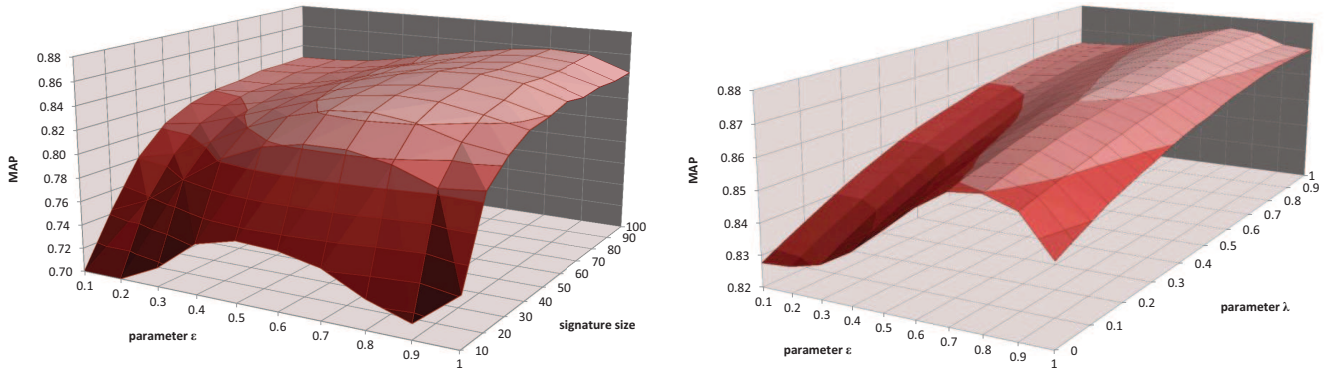
Let us now compare the SMD to the other signature-based approaches. In particular, we compare it to the earth mover's distance (EMD), the perceptually modified Hausdorff distance (PMHD), and the signature quadratic form distance (SQFD). The results are summarized in Table 2, where we report the highest MAP values that we have achieved for the other signature-based approaches and, in order to preserve comparability, the MAP values for the SMD with default parameter $\lambda = 0$ and the nearest neighbor matching strategy m^{NN}. In general, the reported values reveal the same tendency as those shown in Table 1. The highest retrieval performance is achieved when using PCT-based signatures, except the EMD on the Holidays database, where the CSIFT descriptor performs best. The default SMD outperforms the other signature-based approaches on the Holidays database, while it is outperformed by the PMHD on the UKBench database. However, as has been shown in the previous section, the SMD is able to improve the retrieval performance when adapting the parameters appropriately.

We finally analyze the efficiency of the SMD in comparison to the other signature-based approaches. For this purpose, we measured the computation times needed to perform 1 million distance computations on a single-core 3.4 GHz machine. We implemented all signature-based approaches in Java 1.6. On average, the SMD is approximately three times faster than the PMHD, 15 times faster than the SQFD, and 84 times faster than the EMD. By using PCT-based signatures of size 10, the SMD performs 1 million distance computations in 1.4s. This value increases to 82.8s when using PCT-based signatures of size 100. We thus conclude that the SMD is able to outperform other signature-based approaches to content-based image retrieval in terms of efficiency and accuracy.



(a) (b)

Figure 4: Mean average precision values of the SMD on the Holidays database by using the inverse distance ratio matching $m^{i\delta r}$. (a): MAP values as a function of the parameter ε and the signature size (parameter $\lambda = 0$), (b): maximum MAP values over all signature sizes as a function of the parameters ε and λ.



(a) (b)

Figure 5: Mean average precision values of the SMD on the UKBench database by using the inverse distance ratio matching $m^{i\delta r}$. (a): MAP values as a function of the parameter ε and the signature size (parameter $\lambda = 0$), (b): maximum MAP values over all signature sizes as a function of the parameters ε and λ.

5.3 Comparison to BoVW Approaches

We compare the performance of our proposed approach to those of the conventional bag-of-visual-words approaches (BoVW) as well as to recent extensions, namely Hamming embedding (HE), compressed Fisher vectors (FV), vector of locally aggregated descriptors (VLAD), and asymmetric Hamming embedding (AHE). Since our focus lies on investigating the maximum achievable retrieval performance in the context of content-based image retrieval, we report the MAP values of the BoVW approaches obtained by optimizing their parameters and refer to the corresponding research articles for details concerning the parameter selection and optimization of the particular methods.

The retrieval performance results are summarized in Table 3. As can be seen in the table, the classic BoVW approach with a vocabulary size of 20k and 200k visual words is generally outperformed by the extensions of the BoVW model and our signature-based approach. Among the BoVW extensions, the AHE and the FV approaches reach the highest retrieval performance on the Holidays and UKBench databases, respectively. The proposed SMD is able to compete with these approaches. It shows the highest and the second highest retrieval performance on the Holidays and UKBench databases.

Regarding the computation times of recent BoVW approaches, Jain et al. [7] report a value of 1.7s for HE and 2.9s for AHE while Perronnin et al. [16] report a value of 0.4s for FV when searching 1 million images. Jégou et al. [10] report a value of 7.2s for VLAD on a 10 million image dataset which approximately corresponds to 0.72s for 1 million images. All computation times are measured on a single-core CPU. However, since the implementation details are not apparent in the respective research articles, it is infeasible to directly compare the runtimes of the different approaches. We nonetheless report these values in order to give a rough idea on the magnitude of runtime and the applicability for content-based image retrieval tasks.

Summarizing, the performance evaluation shows that the SMD competes with the state-of-the-art approaches to content-based image retrieval by using a simplistic color and texture-based image descriptor. In fact, it outperforms the AHE approach that shows the best results so far [7] on the Holidays database.

Table 3: Retrieval performance of the signature-based approaches and the bag-of-visual-words approaches on the Holidays and UKBench databases.

	Holidays	UKBench	
	MAP	MAP	score
BoVW (20k) (from [9])	0.469	0.752	–
BoVW (200k) (from [9])	0.572	0.771	–
HE (from [9])	0.813	**0.878**	3.42
FV (from [16])	0.735	–	**3.50**
VLAD (from [11])	0.621	–	3.35
AHE (from [7])	0.819	–	–
Our approach	**0.834**	0.876	3.34

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have addressed the problem of content-based image retrieval by means of the signature-based model. To this end, we have outlined two complementary approaches to content-based image retrieval, namely the bag-of-visual-words model and the signature-based model. We investigated different matching strategies and analyzed their suitability to match coincident visual properties between images based on their signatures. We proposed a novel matching strategy, the inverse distance ratio matching, and a novel distance function, the signature matching distance (SMD). Besides the theoretical definition of this distance, we investigated its properties and studied its performance in comparison to both the bag-of-visual-words approaches and the other signature-based approaches on different benchmark image databases.

Our performance evaluation reveals that the SMD is able to outperform other signature-based approaches to content-based image retrieval in terms of efficiency and accuracy. Moreover, the SMD shows a higher performance than the majority of bag-of-visual-words approaches by using a simple color and texture-based image descriptor. We thus conclude, that the SMD is a simple yet effective signature-based approach that is able to compete with the state of the art in content-based image retrieval.

As future work, we plan to investigate signature-based indexing methods in order to apply the SMD to large-scale image retrieval.

Acknowledgments

This work is partially funded by the Excellence Initiative of the German federal and state governments and by DFG grant SE 1039/7-1.

7. REFERENCES

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

[2] C. Beecks, M. S. Uysal, and T. Seidl. Signature Quadratic Form Distance. In *CIVR*, pages 438–445, 2010.

[3] G. J. Burghouts and J.-M. Geusebroek. Performance evaluation of local colour invariants. *CVIU*, 113(1):48–62, 2009.

[4] M. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, 2002.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008.

[6] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493, 1998.

[7] M. Jain, H. Jégou, and P. Gros. Asymmetric hamming embedding: taking the best of our bits for large scale image search. In *ACM Multimedia*, pages 1441–1444, 2011.

[8] H. Jégou, M. Douze, and C. Schmid. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In *ECCV (1)*, pages 304–317, 2008.

[9] H. Jégou, M. Douze, and C. Schmid. Improving Bag-of-Features for Large Scale Image Search. *IJCV*, 87:316–336, 2010.

[10] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010.

[11] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE TPAMI*, 2011.

[12] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004.

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

[14] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR (2)*, pages 2161–2168, 2006.

[15] B. G. Park, K. M. Lee, and S. U. Lee. A New Similarity Measure for Random Signatures: Perceptually Modified Hausdorff Distance. In *ACIVS*, pages 990–1001, 2006.

[16] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In *CVPR*, pages 3384–3391, 2010.

[17] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann. Empirical Evaluation of Dissimilarity Measures for Color and Texture. *CVIU*, 84(1):25 – 43, 2001.

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

[19] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, pages 1470–1477, 2003.

[20] 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(12):1349–1380, 2000.

[21] H. Tamura. Texture features corresponding to visual perception. *IEEE Trans. on Systems, Man, and Cybernetics*, 8(6):460–473, 1978.

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

[23] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, pages 1753–1760, 2008.