Multimed Tools Appl (2014) 71:349–362
DOI 10.1007/s11042-012-1334-3

On stability of signature-based similarity measures for content-based image retrieval

Christian Beecks · Steffen Kirchhoff · Thomas Seidl

Published online: 7 February 2013

Abstract Retrieving similar images from large image databases is a challenging task for today's content-based retrieval systems. Aiming at high retrieval performance, these systems frequently capture the user's notion of similarity through expressive image models and adaptive similarity measures. On the query side, image models can significantly differ in quality compared to those stored on the database side. Thus, similarity measures have to be robust against these individual quality changes in order to maintain high retrieval performance. In this paper, we investigate the robustness of the family of signature-based similarity measures in the context of content-based image retrieval. To this end, we introduce the generic concept of *average precision stability*, which measures the stability of a similarity measure with respect to changes in quality between the query and database side. In addition to the mathematical definition of average precision stability, we include a performance evaluation of the major signature-based similarity measures focusing on their stability with respect to querying image databases by examples of varying quality. Our performance evaluation on recent benchmark image databases reveals that the highest retrieval performance does not necessarily coincide with the highest stability.

Keywords Content-based image retrieval · Feature signature · Distance-based similarity measure · Evaluation measure · Average precision stability

This paper is an extended version of a previous paper by Beecks and Seidl [3].

C. Beecks (✉) · S. Kirchhoff · T. Seidl
Data Management and Data Exploration Group, RWTH Aachen University, Aachen, Germany
e-mail: beecks@cs.rwth-aachen.de

S. Kirchhoff
e-mail: kirchhoff@cs.rwth-aachen.de

T. Seidl
e-mail: seidl@cs.rwth-aachen.de

1 Introduction

Modeling image contents for the purpose of content-based image retrieval [5, 22] is a challenging task. While the computational effort spent for extracting and generating expressive image models is nearly unrestricted on the database side, the effort spent on the query side is often restricted since users usually demand the retrieval system to answer their queries in real-time. As a consequence, the time for the extraction of complex local feature descriptors and for the generation of complex image models has to be kept short, in particular for query images that have not been processed by the retrieval system. This inevitably leads to a quality gap between the query side and the database side. Image models that appear on the query side can significantly differ in quality compared to those stored in the multimedia database. Thus, the similarity measures of the retrieval systems have to be robust against image models of varying qualities as well as capable of processing such models efficiently.

A prominent family of similarity measures that is inherently able to cope with different qualities of image models is that of distance-based similarity measures. Based on a solid mathematical definition, distance-based similarity measures allow domain experts to model their notion of similarity even if the similarity model has to be subjected to different quality restrictions. At the same time, they allow database experts to design efficient query processing approaches including index structures, such as metric access methods [4, 26]. Although the performance of similarity measures for different types of image models is investigated in various studies [1, 7, 20], none of them addresses the issue of query-side-dependent quality restrictions. They all assume the quality of the image model on the query side to be the same as on the database side. For this reason, we study the *stability* of the most generic class of distance-based similarity measures, namely the *signature-based similarity measures*, in the context of content-based image retrieval. We investigate the retrieval performance with respect to stability against varying image model quality of the *Perceptually Modified Hausdorff Distance* [18], the *Earth Mover's Distance* [19], the *Weighted Correlation Distance* [13], and the *Signature Quadratic Form Distance* [2]. To this end, we first introduce the generic concept of *average precision stability*, which measures the stability of a similarity measure with respect to changes in quality between the query and the database side. We then provide a performance evaluation of the aforementioned signature-based similarity measures focusing on their stability with respect to querying image databases by examples of varying quality. Without loss of generality, the proposed average precision stability is generic enough to be used in conjunction with other applicable evaluation measures.

This paper is structured as follows. We describe the feature signature model in Section 2, which comprises feature signatures and signature-based similarity measures. Then, in Section 3, we outline existing evaluation measures, which can be used within the average precision stability measure we propose in Section 4. We evaluate the stability of signature-based similarity measures on different benchmark image databases in Section 5, before we conclude our paper in Section 6.

2 Feature signature model

A common way to make images accessible consists in describing their contents by feature distributions over a feature space. While many similarity models that are designed against the background of visual object recognition tasks rely on complex unaggregated local features, similarity models for the purpose of content-based image retrieval frequently aggregate individual feature distributions in order to obtain more compact and robust content representations.

In general, the modeling of image contents follows a two-stage approach. First, local features are extracted, for instance SIFT [14] descriptors at some salient points [16, 24]. Second, these features are aggregated into a more compact representation. One prominent way of aggregating and comparing the extracted local features is by means of the *bag-of-visual-words* [21] approach. Based on a predetermined *visual vocabulary*, the extracted local features are assigned to *visual words*. The frequency of these visual words is then used in order to define similarity between images. Although this approach provides high retrieval performance, it is limited in flexibility due to the static visual vocabulary. In fact, all images have to be represented by the same visual words. Moreover, the presence of the visual words of the database side has to be ensured on the query side in order to compute an image content representation that is compatible with the database.

An alternative approach to model image contents is epitomized by the *feature signature model*. Local features are extracted and quantized for each image individually by means of an image-specific visual vocabulary, namely the *feature signature* [19]. A feature signature X quantizes a feature space $\mathbb{F}$ by a finite set of representatives $\mathrm{R}_X \subset \mathbb{F}$, where each representative is additionally assigned to a certain positive weight by a weighting function $w_X : \mathrm{R}_X \to \mathbb{R}^{\geq 0}$. Mathematically, a feature signature X can be defined as the graph of its weighting function w_X:

$$X = \{(x, w_X(x)) | x \in \mathrm{R}_X\}.$$

Furthermore, the set of *all* feature signatures $\mathbb{S}$ can be defined via the set of *all* weighting functions $w : \mathrm{R} \to \mathbb{R}$ with a finite set of representatives R, i.e.: $\mathbb{S} = \bigcup_{\mathrm{R} \subset \mathbb{F}, |\mathrm{R}| < \infty} \mathbb{R}^{\mathrm{R}}$. Given an image, its feature signature X can be computed by clustering its local features and defining the representatives R_X of the feature signature by the centroids of the clusters. The weighting function can be defined through the corresponding cluster sizes. In this way, the representatives and weights of a feature signature correspond to visual words and their frequencies of an image-specific visual vocabulary.

In Fig. 1, we depict three example images from the *MIR Flickr* database [8] together with their feature signatures. These feature signatures were generated by mapping randomly selected image pixels into a seven-dimensional feature space $(L, a, b, x, y, \chi, \eta) \in \mathbb{F} = \mathbb{R}^7$ that comprises color (L, a, b), position (x, y), contrast χ, and coarseness η. The extracted seven-dimensional features are clustered by the k-means clustering algorithm in order to obtain the feature signatures. As can be seen in the figure, the higher the number of centroids, which are depicted as circles in the corresponding color, the better the visual content approximation, and vice

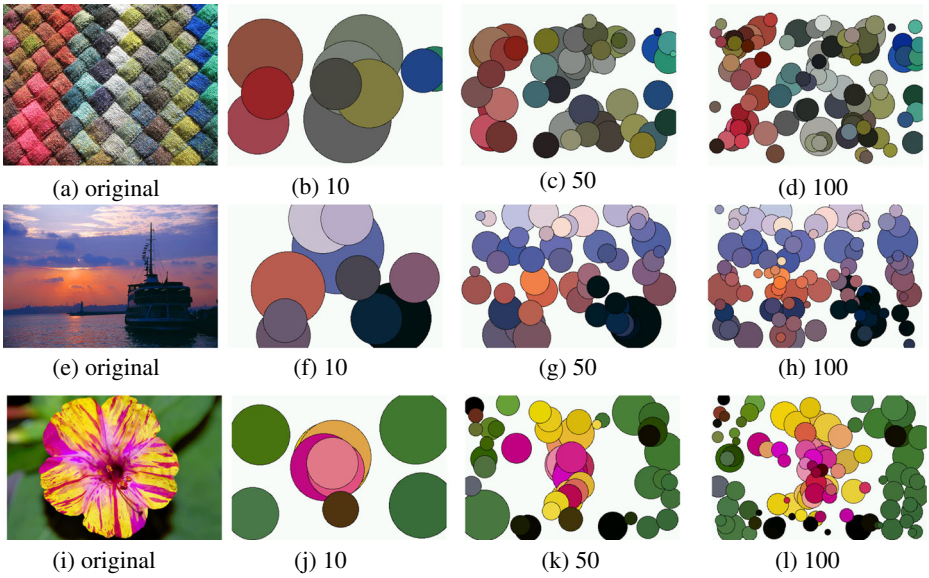


Fig. 1 Three example images from the *MIR Flickr* database [8] and their corresponding feature signatures over a feature space comprising position, color, and texture information. The number of representatives, i.e., the centroids, is depicted accordingly

versa. While a small number of centroids only provides a coarse approximation of the original image, a large number of centroids may help to assign individual centroids to the corresponding parts in the images.

Based on the feature signature representations, a distance function is applied in order to determine a similarity value between the corresponding images. For this purpose, the distances between feature signatures utilize a so-called *ground distance* $\delta : \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R}$ to measure the distance between two representatives of the feature signatures. An overview of applicable distances to feature signatures can be found, for instance, in the work of Beecks et al. [1]. We summarize the major distances in the remainder of this section.

2.1 Earth Mover's Distance

The *Earth Mover's Distance* (EMD) [19] is a *transformation-based* approach measuring the costs of transforming two feature signatures into another. Given two feature signatures $X, Y \in \mathbb{S}$ and a ground distance δ, the Earth Mover's Distance EMD_δ between X and Y is defined as a minimum cost flow over all possible flows $[F \in \mathbb{R}^{|\mathrm{R}_X| \times |\mathrm{R}_Y|}]$ between two elements $x, y \in \mathrm{R}_X \cup \mathrm{R}_Y$ as:

$$\mathrm{EMD}_\delta(X, Y) = \min_F \left\{ \frac{\sum_{x \in \mathrm{R}_X} \sum_{y \in \mathrm{R}_Y} f_{xy} \cdot \delta(x, y)}{\min\left\{ \sum_{x \in \mathrm{R}_X} w_X(x), \sum_{y \in \mathrm{R}_Y} w_Y(y) \right\}} \right\},$$

subject to the constraints $\forall x, y : f_{xy} \geq 0$, $\forall x \in \mathrm{R}_X : \sum_{y \in \mathrm{R}_Y} f_{xy} \leq w_X(x)$, $\forall y \in \mathrm{R}_Y : \sum_{x \in \mathrm{R}_X} f_{xy} \leq w_Y(y)$, and $\sum_{x \in \mathrm{R}_X} \sum_{y \in \mathrm{R}_Y} f_{xy} = \min\{\sum_{x \in \mathrm{R}_X} w_X(x), \sum_{y \in \mathrm{R}_Y} w_Y(y)\}$.

2.2 Perceptually Modified Hausdorff Distance

The *Perceptually Modified Hausdorff Distance* (PMHD) [18] is a *matching-based* approach. In general, a *matching* between two feature signatures $X, Y \in \mathbb{S}$ can be defined as $\mathrm{m}_{X\to Y} = \{(x, \pi_{X\to Y}(x)) | \forall x \in \mathrm{R}_X\}$, where the *matching function* $\pi_{X\to Y} : \mathrm{R}_X \to \mathrm{R}_Y$ maps each representative $x \in \mathrm{R}_X$ to one representative $y \in \mathrm{R}_Y$. Additionally, a matching can be evaluated by a cost function $\mathrm{c} : 2^{\mathrm{R}_X \times \mathrm{R}_Y} \to \mathbb{R}^{\geq 0}$. Based on a matching, a cost function and a ground distance δ, the Perceptually Modified Hausdorff Distance PMHD_δ between X and Y is defined as:

$$\mathrm{PMHD}_\delta(X, Y) = \max\{\mathrm{c}(\mathrm{m}_{X\to Y}), \mathrm{c}(\mathrm{m}_{Y\to X})\},$$

where the matching $\mathrm{m}_{X\to Y}$ is defined by the graph of the matching function $\pi_{X\to Y}(x) = \mathrm{argmin}_{y\in \mathrm{R}_Y}\left\{\frac{\delta(x,y)}{\min\{w_X(x), w_Y(y)\}}\right\}$, and the costs c of the matching are given by $\mathrm{c}(\mathrm{m}_{X\to Y}) = \sum_{(x,y)\in \mathrm{m}_{X\to Y}} \frac{w_X(x)}{\sum_{(x,y)\in \mathrm{m}_{X\to Y}} w_X(x)} \cdot \frac{\delta(x,y)}{\min\{w_X(x), w_Y(y)\}}$. The matching $\mathrm{m}_{Y\to X}$ is defined analogously.

2.3 Signature Quadratic Form Distance

The *Signature Quadratic Form Distance* (SQFD) [2] is a *correlation-based* approach. In contrast to the two approaches described above, it uses a symmetric *similarity function* $\mathrm{s} : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ in order to express how similar two representatives of the feature signatures are. Further, given a similarity function s, the *weighted similarity correlation* between two feature signatures X and Y is defined 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 SQFD_s between X and Y is defined as:

$$\mathrm{SQFD}_\mathrm{s}(X, Y) = \sqrt{X \cdot_\mathrm{s} X - X \cdot_\mathrm{s} Y - Y \cdot_\mathrm{s} X + Y \cdot_\mathrm{s} Y}.$$

2.4 Weighted Correlation Distance

Another correlation-based approach is the *Weighted Correlation Distance* (WCD) [13]. It utilizes the weighted similarity correlation that is defined by the specific weighting function $\mathrm{w} : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ with maximum cluster radius $\mathrm{R} \in \mathbb{R}$ as $\mathrm{w}(x, y) = 1 - \frac{3\cdot\delta(x,y)}{4\cdot \mathrm{R}} + \frac{\delta(x,y)^3}{16\cdot \mathrm{R}^3}$ if $0 \leq \frac{\delta(x,y)}{\mathrm{R}} \leq 2$ and $\mathrm{w}(x, y) = 0$ otherwise. The Weighted Correlation Distance WCD between X and Y is then defined as:

$$\mathrm{WCD}(X, Y) = 1 - \frac{X \cdot_\mathrm{w} Y}{\sqrt{X \cdot_\mathrm{w} X} \cdot \sqrt{Y \cdot_\mathrm{w} Y}}.$$

Given the combination of feature signatures and signature-based similarity measures, several studies [1, 7, 20] have considered the fundamental question of the highest retrieval performance of different similarity models. These studies followed the assumption that the quality of the image model is the same on both query and database side. Since the quality of the feature signatures on the query side is in general unpredictable, the focus of this paper lies in the following additional question: which signature-based similarity measure is the most robust one provided that the quality on the query side differs from that on the database side. The answer of this question leads to the generic concept of *average precision stability*, which we

introduce in Section 4. Prior to this, we first describe existing evaluation measures that are commonly used within the content-based retrieval community.

3 Evaluation measures

In general, evaluating a similarity measure is done by querying an image collection and analyzing the results. For this purpose, the image collection is sorted in descending order according to their similarity regarding the query image, i.e., the retrieval system computes a *ranking* of the database, and each image is assigned a class label. The class labels are provided by the *ground truth* of the image collection and define the *relevancy* of each image with respect to the query image. A good overview of measuring the effectiveness of a retrieval system and a broad introduction to several evaluation measures can be found, for instance, in the book of Manning et al. [15].

In fact, many evaluation measures are based on *precision* and *recall* values—first used by Kent et al. [12]—which reflect the fraction of retrieved images that are relevant and the fraction of relevant images that are retrieved [15], respectively. Thus, a high precision value indicates that most of the retrieved images are relevant while a high recall value indicates that most of the relevant images are retrieved. These values can be computed for each retrieved image within the ranking and can then be visualized by the so-called *precision and recall curve*. A frequently encountered aggregation of multiple precision and recall curves is the *Mean Average Precision* value, which approximates the average area under the curves [15]. Other evaluation measures are the *F-Measure* [25], which is the weighted harmonic mean of precision and recall [15], or the *Normalized Discounted Cumulative Gain* [10], which measures the usefulness of multiple rankings.

Summarizing, the aforementioned evaluation measures judge the retrieval performance according to a single ranking or multiple rankings. Although they are frequently used throughout the research area of content-based retrieval, see for instance the performance evaluations for content-based image retrieval [1, 7, 20], they miss the ability to express the variance of a measured quantity. For instance, measuring the same Mean Average Precision values for two different similarity measures does not necessarily mean that both similarity measures show the same retrieval performance. One similarity measure can show a higher variance than the other one, which is, in this example, not reflected within the Mean Average Precision values.

In order to counteract this issue, we propose to include the stability of a similarity measure into the evaluation of the retrieval performance. Our approach that takes into account the stability is described in the next section.

4 Stability of a similarity measure

As mentioned above, our interests lie in evaluating the *stability* of signature-based similarity measures in the context of content-based image retrieval. In particular, the stability of a signature-based similarity measure with respect to the changes in quality between the query side and the database side offers further insight into the behavior of those measures and will thus help to guide further research and developments.

A general concept to model these quality changes between the query and the database side is that of *query modifying transformations*. They provide a solid mathematical means of reflecting the general discrepancy between the image models generated on the query side and those stored in the image database. Without loss of generality, we assume that the modifications of the image models are only done on the query side. Further, we focus on the evaluation measure of Mean Average Precision in the remainder of this paper, as this is the de facto standard in content-based image retrieval. Note that Mean Average Precision can be replaced by any other evaluation measure when desired. However, by using Mean Average Precision (MAP) as evaluation measure, we denote our resulting stability measure as *Average Precision Stability* (APS). It is generally defined for a similarity measure δ over an image database $\mathcal{DB}$, a set of queries Q, and a set of query modifying transformations Φ as follows.

Definition 1 Average Precision Stability (APS). Given a similarity measure δ, a database $\mathcal{DB}$, a set of queries $Q = \{q_1, \ldots, q_l\}$, and a set of query modifying transformations $\Phi = \{\phi_1, \ldots, \phi_m\}$, the Average Precision Stability (APS) is then defined as:

$$\mathrm{APS}_{\Phi}(Q, \delta, \mathcal{DB}) = \frac{\mathrm{E}[\mathbb{M}]}{1 + \sigma_{\mathbb{M}}},$$

where $\mathbb{M}$ denotes the distribution of Mean Average Precision values with respect to the query modifying transformations $\Phi = \{\phi_1, \ldots, \phi_m\}$ applied to each query contained in the set of queries Q, i.e. $\mathbb{M} = \bigcup_{i=1}^{m} \{\mathrm{MAP}(\{\phi_i(q_1), \ldots, \phi_i(q_l)\}, \delta, \mathcal{DB})\}$. $\mathrm{E}[\mathbb{M}]$ and $\sigma_{\mathbb{M}}$ denote the expected value and standard deviation, respectively.

According to Definition 1, the *Average Precision Stability* is defined as the expected Mean Average Precision value divided by the standard deviation of those Mean Average Precision values with respect to a set of query modifying transformations. In this way, it reflects the stability of a similarity measure as follows. In case the similarity measure is invariant against the query modifying transformations, the *Average Precision Stability* becomes the expected Mean Average Precision value. Otherwise, the *Average Precision Stability* decreases with varying Mean Average Precision values. As can be seen in the definition, the proposed *Average Precision Stability* generalizes the Mean Average Precision measure by including the standard deviation of the Mean Average Precision values. Consequently, it is also bounded between 0 and 1.

In general, this concept of the stability of a similarity measure can be extended to any other evaluation measure, for instance the *F-Measure* or the *Normalized Discounted Cumulative Gain*, by replacing the evaluation measure appropriately. It is thus flexible to fit individual user and system requirements when evaluating the retrieval performance of content-based multimedia retrieval systems. However, since Mean Average Precision is a frequently encountered evaluation measure in the area of content-based multimedia retrieval, we provide an *Average Precision Stability* evaluation study of signature-based similarity measures for the purpose of content-based image retrieval in the following section.

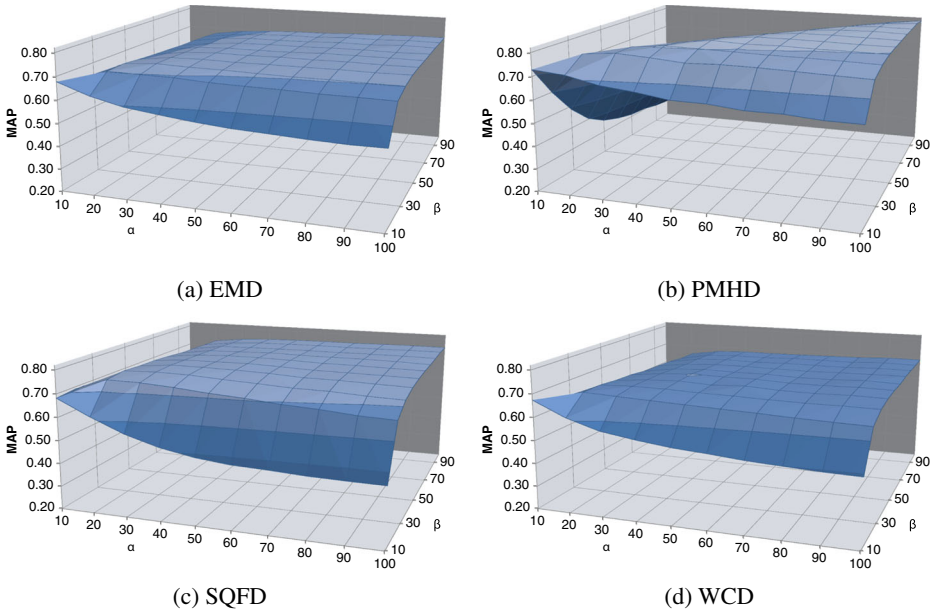
5 Experimental evaluation

In this section, we study the stability of the signature-based similarity measures presented in Section 2 in the context of content-based image retrieval. For this purpose, we used the Holidays [11], UKBench [17], and Copydays [6] image databases, all providing a solid ground truth reflecting photometric and geometric transformations in order to benchmark content-based 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. Within these two databases, the first image of each object or scene serves as query object. The Copydays database comprises 157 images which have been cropped by 50 %. The cropped images serve as query objects for the original images of the Copydays database that is enlarged with 10.000 additional images from the MIR Flickr database [9].

Based on these image databases, we generated feature signatures by extracting local feature descriptors and by clustering them with the k-means algorithm. We extracted a low-dimensional descriptor denoted by PCT [1], which describes the relative spatial information of a pixel, its CIELAB color value, and its first and second Tamura texture features [23], the coarseness and contrast. The PCT descriptor was extracted with a random sampling of 40,000 pixels per image. After having extracted the local feature descriptors, we applied the k-means clustering algorithm to generate multiple feature signatures per image by varying the feature signature size between 10 and 100.

We first investigated the retrieval performance in terms of Mean Average Precision (MAP) for different changes in cardinality between the query signature size $\alpha \in [10, \ldots, 100]$ and the database signature size $\beta \in [10, \ldots, 100]$ in order to evaluate the stability with respect to the most natural modification regardless of any specific local features. The resulting MAP values for the Holidays database are reported in Fig. 2 for the Earth Mover's Distance (EMD), the Perceptually Modified Hausdorff Distance (PMHD), the Signature Quadratic Form Distance (SQFD), and the Weighted Correlation Distance (WCD), where we used the Euclidean distance L_2 as ground distance. While the Earth Mover's Distance and the Perceptually Modified Hausdorff Distance are free of any additional parameter, the Weighted Correlation Distance requires the definition of an appropriate maximum cluster radius $R \in \mathbb{R}$, see Section 2. Based on the applied k-means clustering, we adapt the maximum cluster radius R to the average ground distance between the representatives of each feature signature of the current database throughout our experimental evaluation, as this shows the highest retrieval performance. Similarly, we adapt the parameter of the Gaussian similarity function used within the Signature Quadratic Form Distance to the reciprocal average ground distance between the representatives of each feature signature of the current database minus a constant value of 1.5. In this way, both the Weighted Correlation Distance and the Signature Quadratic form Distance are adjusted to the individual characteristics of the corresponding image databases.

Figure 2 reveals, that all of the signature-based similarity measures are able to achieve a retrieval performance in terms of Mean Average Precision of greater than 70 % on the Holidays database. In particular, the highest Mean Average Precision



(a) EMD (b) PMHD

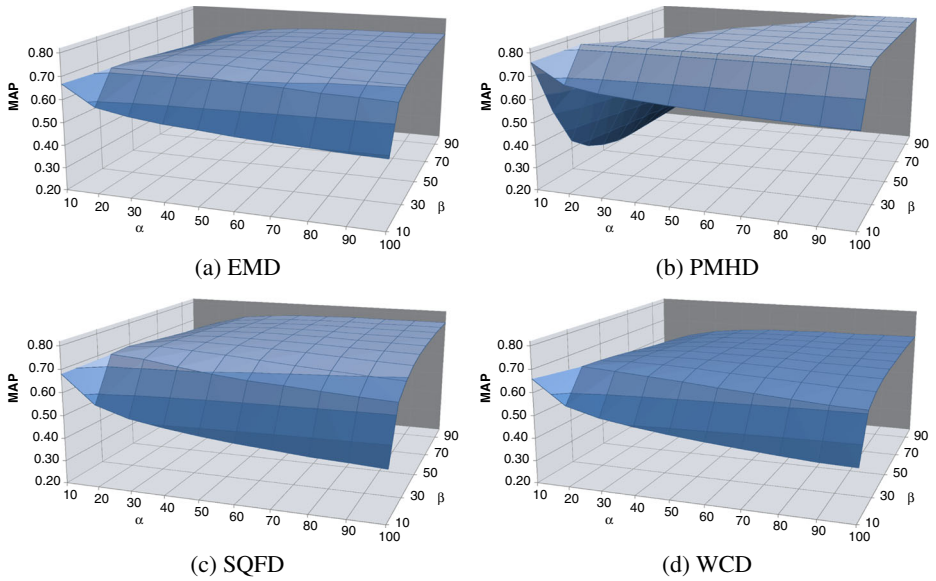
(c) SQFD (d) WCD

Fig. 2 Mean average precision (MAP) values on the Holidays database [11] as a function of the query signature size $\alpha \in [10, \ldots, 100]$ and the database signature size $\beta \in [10, \ldots, 100]$ for the following distances: **a** Earth Mover's Distance (EMD), **b** Perceptually Modified Hausdorff Distance (PMHD), **c** Signature Quadratic Form Distance (SQFD), and **d** Weighted Correlation Distance (WCD)

value of 0.813 is reached by using the Perceptually Modified Hausdorff Distance in combination with a query signature size of $\alpha = 90$ and a database signature size of $\beta = 70$. This Mean Average Precision value is followed by a value of 0.761 using the Signature Quadratic Form Distance with lower signature sizes of $\alpha = 40$ and $\beta = 60$. The Earth Mover's Distance reaches a Mean Average Precision value of 0.722 with comparatively high signature sizes of $\alpha = 90$ and $\beta = 80$, while the Weighted Correlation Distance reaches a value of 0.701 with signature sizes of $\alpha = 40$ and $\beta = 50$.

The results of the UKBench database, reported in Fig. 3, show the same tendency as those of the Holidays database. By using a query signature size of $\alpha = 100$ and a database signature size of $\beta = 80$ the Perceptually Modified Hausdorff Distance reaches the highest Mean Average Precision value of 0.875. The second-highest value of 0.766 is reached by the Signature Quadratic Form Distance with the feature signature sizes of $\alpha = 50$ and $\beta = 60$. The Earth Mover's Distance requires the feature signature sizes of $\alpha = 60$ and $\beta = 50$ to reach a value of 0.742, while the Weighted Correlation Distance stays at a value of 0.695 with the feature signature sizes of $\alpha = 40$ and $\beta = 50$.

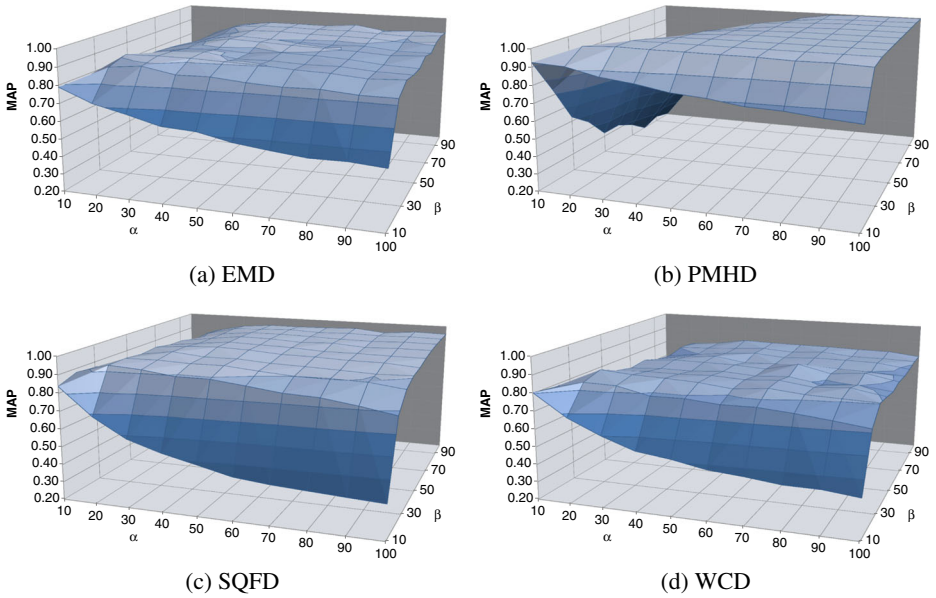
Finally, the results of the Copydays database are shown in Fig. 4. Similar to both databases before, the Perceptually Modified Hausdorff Distance reaches the highest Mean Average Precision value of 1.0 when using a query signature size of $\alpha = 30$ and a database signature size of $\beta = 20$. The Signature Quadratic Form Distance



(a) EMD (b) PMHD

(c) SQFD (d) WCD

Fig. 3 Mean average precision (MAP) values on the UKBench database [17] as a function of the query signature size $\alpha \in [10, \ldots, 100]$ and the database signature size $\beta \in [10, \ldots, 100]$ for the following distances: **a** Earth Mover's Distance (EMD), **b** Perceptually Modified Hausdorff Distance (PMHD), **c** Signature Quadratic Form Distance (SQFD), and **d** Weighted Correlation Distance (WCD)



(a) EMD (b) PMHD

(c) SQFD (d) WCD

Fig. 4 Mean average precision (MAP) values on the Copydays database [6] as a function of the query signature size $\alpha \in [10, \ldots, 100]$ and the database signature size $\beta \in [10, \ldots, 100]$ for the following distances: **a** Earth Mover's Distance (EMD), **b** Perceptually Modified Hausdorff Distance (PMHD), **c** Signature Quadratic Form Distance (SQFD), and **d** Weighted Correlation Distance (WCD)

Table 1 Overview of the minimum, average, and maximum Mean Average Precision values for the Holidays, UKBench, and Copydays image databases

Database	MAP	EMD	PMHD	SQFD	WCD
Holidays	max	0.722	0.813	0.761	0.701
	avg	0.683	0.705	0.712	0.662
	min	0.533	0.276	0.458	0.492
UKBench	max	0.742	0.875	0.766	0.695
	avg	0.688	0.707	0.704	0.648
	min	0.490	0.153	0.422	0.428
Copydays	max	0.925	1.000	0.958	0.868
	avg	0.842	0.873	0.870	0.775
	min	0.532	0.195	0.394	0.424

shows the second-highest value of 0.958 when using a query signature size of $\alpha = 40$ and a database signature size of $\beta = 70$. The Earth Mover's Distance requires the feature signature sizes of $\alpha = 70$ and $\beta = 100$ to reach a value of 0.925, while the Weighted Correlation Distance stays at a value of 0.868 with the feature signature sizes of $\alpha = 20$ and $\beta = 30$. The minimum, average, and maximum Mean Average Precision values for the combination of all image databases and all signature-based similarity measures are summarized in Table 1.

In general, the aforementioned signature-based similarity measures show a similar behavior on all image databases. First, it can generally be observed that the correlation-based approaches, i.e., the Signature Quadratic Form Distance and the Weighted Correlation Distance, require smaller feature signatures in order to achieve high retrieval performance in comparison to the transformation-based and matching-based approaches, i.e., the Earth Mover's Distance and the Perceptually Modified Hausdorff Distance. Second, the retrieval performance deteriorates in case the query signatures or the database signatures are of low cardinality. Third, in particular the Perceptually Modified Hausdorff Distance shows a significant loss in retrieval performance when the cardinality of the query signatures is larger than that of the database signatures, as can be seen in Figs. 2b, 3b, and 4b. Finally, all signature-based similarity measures show a comparatively stable plane of high Mean Average Precision values for a large number of cardinality changes between the query and the database side. Nevertheless, the fluctuations in the marginal areas affect the stability of the corresponding similarity measures, as can be seen in Table 2, where we finally report the Average Precision Stability.

Summarizing, it can be seen in the Table 2 that the Signature Quadratic Form Distance shows the highest stability with respect to changes in cardinality. It reaches the Average Precision Stability of 0.661, 0.645, and 0.763 on the Holidays, UKBench, and Copydays databases, respectively. The second-highest stability is reached by the Earth Mover's Distance, followed by the Weighted Correlation Distance and, finally, the Perceptually Modified Hausdorff Distance. This behavior is reasoned in the correlation-based and transformation-based nature of the first mentioned similarity measures, i.e. the Signature Quadratic Form Distance, the Weighted Correlation

Table 2 Average Precision Stability (APS) for the Holidays, UKBench, and Copydays image databases

Database	EMD	PMHD	SQFD	WCD
Holidays	0.649	0.622	0.661	0.629
UKBench	0.640	0.588	0.645	0.607
Copydays	0.763	0.722	0.763	0.707

Distance, and the Earth Mover's Distance. On the one hand, by taking into account the complete similarity structure of the feature signatures, these similarity measures are more robust to changes in quality compared to the matching-based Perceptually Modified Hausdorff Distance. On the other hand, the latter provides higher absolute retrieval performance.

We conclude that the highest retrieval performance does not necessarily coincide with the highest stability. Thus, a hybrid approach combining both correlation-based and matching-based nature might provide a signature-based similarity measure that is able to reduce the gap between high absolute retrieval performance and high stability.

6 Summary and conclusions

We investigated the stability of the major signature-based similarity measures with respect to quality changes between the query and the database side. For this purpose, we first described the feature signature model, which comprises feature signatures and signature-based similarity measures. We then outlined existing evaluation measures and pointed out their missing ability to express the variance of a measured quantity. In order to counteract this issue, we defined the *Average Precision Stability* by means of the concept of query modifying transformations. Based on this measure, we finally evaluated the stability of the signature-based similarity measures with regard to changes in cardinality between the query and the database signatures.

Our performance evaluation on three recent benchmark image databases including photometric and geometric modifications reveals a gap between the highest retrieval performance and the highest stability. While the first is reached by the matching-based Perceptually Modified Hausdorff Distance, the latter is reached by the correlation-based Signature Quadratic Form Distance.

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

References

1. Beecks C, Uysal MS, Seidl T (2010) A comparative study of similarity measures for content-based multimedia retrieval. In: Proc. IEEE international conference on multimedia & expo, pp 1552–1557
2. Beecks C, Uysal MS, Seidl T (2010) Signature quadratic form distance. In: Proc. ACM international conference on image and video retrieval, pp 438–445
3. Beecks C, Seidl T (2012) On stability of adaptive similarity measures for content-based image retrieval. In: MMM, pp 346–357
4. Chávez E, Navarro G, Baeza-Yates R, Marroquín JL (2001) Searching in metric spaces. ACM Comput Surv 33(3):273–321. doi:10.1145/502807.502808
5. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):1–60. doi:10.1145/1348246.1348248
6. Douze M, Jegou H, Sandhawalia H, Amsaleg L, Schmid C (2009) Evaluation of gist descriptors for web-scale image search. In: CIVR
7. Hu R, Rüger S, Song D, Liu H, Huang Z (2008) Dissimilarity measures for content-based image retrieval. In: Proc. IEEE international conference on multimedia & expo, pp 1365 –1368. doi:10.1109/ICME.2008.4607697

8. Huiskes MJ, Lew MS (2008) The MIR flickr retrieval evaluation. In: Proc. of the 1st ACM international conference on multimedia information retrieval, pp 39–43
9. Huiskes MJ, Thomee B, Lew MS (2010) New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. In: MIR '10: Proceedings of the 2010 ACM international conference on multimedia information retrieval. ACM, New York, pp 527–536
10. Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20:422–446. doi:10.1145/582415.582418
11. Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geometric consistency for large scale image search. In: ECCV (1), pp 304–317
12. Kent A, Berry MM, Luehrs FU, Perry JW (1955) Machine literature searching VIII. Operational criteria for designing information retrieval systems. Am Doc 6(2):93–101. doi:10.1002/asi.5090060209
13. Leow WK, Li R (2004) The analysis and applications of adaptive-binning color histograms. Comput Vis Image Underst 94(1–3):67–91. doi:10.1016/j.cviu.2003.10.010
14. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110. doi:10.1023/B:VISI.0000029664.99615.94. URL:http://portal.acm.org/citation.cfm?id=993451.996342
15. Manning CD, Raghavan P, Schütze H (2008) Introduction to Information Retrieval. Cambridge University Press, New York
16. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630. doi:10.1109/TPAMI.2005.188
17. Nistér D, Stewénius H (2006) Scalable recognition with a vocabulary tree. In: CVPR (2), pp 2161–2168
18. Park BG, Lee KM, Lee SU (2008) Color-based image retrieval using perceptually modified Hausdorff distance. J Image Video Process 2008:1–10. doi:10.1155/2008/263071
19. Rubner Y, Tomasi C, Guibas LJ (2000) The Earth Mover's Distance as a metric for image retrieval. Int J Comput Vis 40(2):99–121. doi:10.1023/A:1026543900054
20. Rubner Y, Puzicha J, Tomasi C, Buhmann JM (2001) Empirical evaluation of dissimilarity measures for color and texture. Comput Vis Image Underst 84(1):25–43
21. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: IEEE international conference on computer vision, pp 1470–1477
22. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380. doi:10.1109/34.895972
23. Tamura H (1978) Texture features corresponding to visual perception. IEEE Trans Syst Man Cybern 8(6):460–473
24. Tuytelaars T, Mikolajczyk K (2008) Local invariant feature detectors: a survey. Found Trends Comput Graph Vis 3(3):177–280. doi:10.1561/0600000017
25. van Rijsbergen CJ (1979) Information retrieval. Butterworth, Boston
26. Zezula P, Amato G, Dohnal V, Batko M (2005) Similarity search: the metric space approach. Springer, New York

Christian Beecks is a PhD student in computer science in the data management and data exploration group at RWTH Aachen University, Germany. His research interests include efficient and effective

content-based multimedia retrieval and exploration, and adaptive distance-based similarity measures. His current particular research interest is devoted to the Signature Quadratic Form Distance.

Steffen Kirchhoff is a research associate in the data management and data exploration group at RWTH Aachen University, Germany. His research interests include efficient and effective content-based multimedia retrieval and adaptive distance-based similarity measures.

Thomas Seidl is a professor of computer science and head of the data management and data exploration group at RWTH Aachen University, Germany. His research interests include data mining and database technology for multimedia and spatio-temporal databases in engineering, communication, and life science applications. Prof. Seidl received his Diploma (MSc) in 1992 from TU Muenchen and his PhD (1997) and venia legendi (2001) from LMU Muenchen.