On Stability of Adaptive Similarity Measures for Content-Based Image Retrieval

Christian Beecks and Thomas Seidl

Data Management and Data Exploration Group
RWTH Aachen University
Germany
{beecks,seidl}@cs.rwth-aachen.de

Abstract. Retrieving similar images 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, which try to approximate the individual user-dependent notion of similarity as close as possible. As image models appearing on the query side can significantly differ in quality compared to those stored in the multimedia database, similarity measures have to be robust against these individual quality changes in order to maintain high retrieval performance. In order to evaluate the robustness of similarity measures, we introduce the general concept of the *stability of a similarity measure with respect to query modifying transformations* describing the change in quality on the query side. In addition, we include a comparison of the stability of the major state-of-the-art adaptive similarity measures based on different benchmark image databases.

Keywords: content-based image retrieval, feature signature, adaptive similarity measure, evaluation measure, average precision stability.

1 Introduction

Modeling image contents for the purpose of content-based image retrieval [4,21,23,13] 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 limited due to the following reasons: first, users frequently demand the retrieval system to answer their queries as fast as possible, thus including the extraction of complex local feature descriptors is a time consuming task which has to be done quickly or even skipped. Second, users issuing queries in a mobile environment, e.g., by taking a picture with a mobile phone, are often restricted in terms of their devices' energy consumption and bandwidth restrictions. As a consequence, processing images with the aim of generating expressive image models has to be kept short which inevitably leads to a gap of quality between the query side and the database

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 346–357, 2012.

side. Image models appearing on the query side can significantly differ in quality compared to those stored in the multimedia database. Thus, the retrieval system's similarity measure has to be robust against these individual changes.

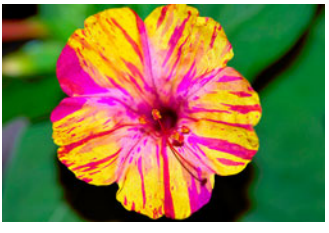
Although the performance of similarity measures for different types of image models is investigated in various studies [2,7,19], none of them addresses the issue of query-side-dependent quality restrictions. They all assume the quality of the image model on the query side is the same as that on the database side. For this reason, we study the *stability* of adaptive similarity measures, namely the *Hausdorff Distance* [9], *Perceptually Modified Hausdorff Distance* [18], *Earth Mover's Distance* [20], *Weighted Correlation Distance* [12], and *Signature Quadratic Form Distance* [1,3], in the context of content-based image retrieval. To this end, we first introduce the general concept of the *stability of a similarity measure with respect to query modifying transformations*, and we then evaluate this stability on different benchmark image databases by using *Mean Average Precision* [15] as a running example.

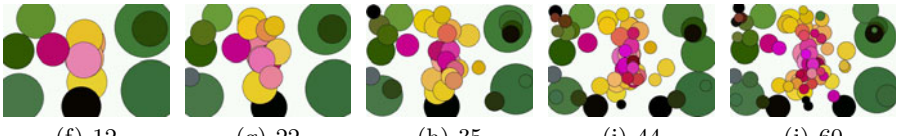
The structure of this paper is as follows: in Section 2, we describe feature signatures as flexible image models and list adaptive similarity measures applicable to such models. In Section 3, we outline existing evaluation measures which can be used within our proposed stability measure. In Section 4, we introduce the general concept of the *stability of a similarity measure* and explain the differences to existing evaluation measures. We evaluate the stability of the adaptive similarity measures on different benchmark image databases in Section 5, before we conclude our paper with an outlook on future work in Section 6.

2 Modeling and Comparing Image Contents through Feature Signatures and Adaptive Similarity Measures

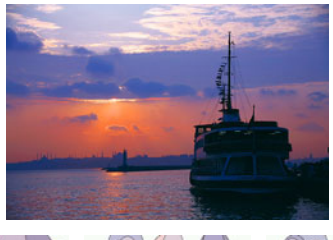
Describing the content of an image by its feature distribution over a feature space is a common way to make images accessible. While many similarity models, which cope with visual object recognition tasks such as near-duplicate detection, rely on complex unaggregated local features, similarity models for the purpose of content-based multimedia retrieval frequently aggregate individual feature distributions in order to obtain more compact and robust content representations. In general, modeling image content follows two steps: 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 the *bag-of-visual-words* [22] approach. Based on a predetermined *visual vocabulary*, the extracted local features are assigned to the *visual words* of that specific visual vocabulary. The similarity between images is then defined through a distance between the visual word frequencies, stored in form of a vector. 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, resulting in a sparse high-dimensional vector representation. Moreover, the availability of the database's visual vocabulary has to be ensured on the query side in order to compute the content representation of an image.

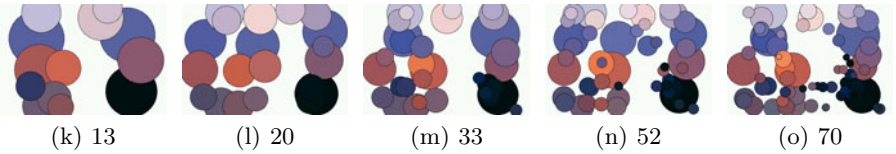
(a) 8 (b) 26 (c) 48 (d) 72 (e) 110





(f) 12 (g) 22 (h) 35 (i) 44 (j) 60





(k) 13 (l) 20 (m) 33 (n) 52 (o) 70

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. centroids, is depicted accordingly.

An alternative of aggregating and comparing the images' local features is by making use of *adaptive similarity measures* [2], which are independent of a visual vocabulary. They allow to compare images whose local features are extracted and aggregated individually. Formally, each image $\mathcal{I}$ is mapped to a set of local features $f_1, \ldots, f_n \in \mathbb{F}$ within a feature space $\mathbb{F}$. Subsequently, these

features are partitioned by a partitioning $\mathcal{P} = \{P_1, \ldots, P_k\}$ where each feature f_i is assigned to its nearest partition. As a result, each partition is represented by a representative $r_i \in \mathbb{F}$ and a weight $w_i \in \mathbb{R}^{\geq 0}$ which form the components of a *feature signature* S as follows:

$$S = \{\langle r_i, w_i \rangle | r_i \in \mathbb{F} \wedge w_i \in \mathbb{R}^{\geq 0}\}_{i=1}^{k}.$$

Frequently, the partitioning $\mathcal{P}$ is obtained by the k-means clustering algorithm: the representatives r_i are the centroids of each cluster P_i with weights w_i denoting the relative frequencies, i.e. $r_i = \sum_{f \in P_i} \frac{f}{|P_i|}$ and $w_i = \frac{|P_i|}{\sum_i |P_i|}$.

In Figure 1, we depict three example images and their feature signatures which were generated by mapping randomly selected image pixels into a seven-dimensional feature space $(L, a, b, x, y, \chi, \eta) \in \mathbb{F} = \mathbb{R}^7$ comprising color (L, a, b), position (x, y), contrast χ, and coarseness η information. The extracted seven-dimensional features are clustered by an adaptive variant of the k-means clustering algorithm [12] in order to obtain the feature signatures. Thus, the number of centroids is determined dynamically and controlled by the number of selected image pixels. 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 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. Given these examples, the question arises; which image model, i.e. feature signature, provides the highest retrieval performance? Furthermore, as the quality of a feature signature on the query side is frequently unpredictable, another question arises; which adaptive similarity measure is the most robust one? In particular the evaluation of the latter, the *robustness* or *stability*, is the focus of this paper. Therefore, we introduce the general concept of a *similarity measure's stability with respect to query modifying transformations* in Section 4, after describing existing evaluation measures, which can be used within our proposed stability measure, in the next section.

3 Evaluation Measures

In general, evaluating a similarity measure is done by querying an image collection and analyzing the results. For this purpose, the images are 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 retrieval systems' effectiveness 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. [11] – 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 many relevant images have been retrieved while a high recall value indicates that the complete amount of relevant images is reached by the retrieved images. 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.

To sum up, 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 [2,7,19], they miss the ability to express the variance of a measured value. For example, measuring the same Mean Average Precision values twice 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 into the evaluation of the retrieval performance. As we are focusing on the retrieval performance of adaptive similarity measures for content-based image retrieval, we show how to evaluate the stability of a similarity measure by making use of conventional Mean Average Precision values in the next section.

4 Stability of a Similarity Measure

As mentioned above, we are interested in evaluating the *stability* of adaptive similarity measures in the context of content-based image retrieval with respect to query modifying transformations, which has not been investigated in previous studies [2,7,19] so far. This will provide further insight into the behavior of adaptive similarity measures and will thus help to guide further research and developments.

In order to generally define the stability of a similarity measure, we combine existing evaluation measures, as described in the previous section, with query modifying transformations. These transformations reflect 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 make use of Mean Average Precision as evaluation measure in the remainder of this paper. This evaluation measure can be replaced with any other evaluation measure where appropriate. However, by using Mean Average Precision (MAP) as evaluation measure, we denote our resulting stability measure as *Average Precision*

Stability (APS). It is defined for a similarity measure δ over a database $\mathcal{DB}$ storing the images, 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:*

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

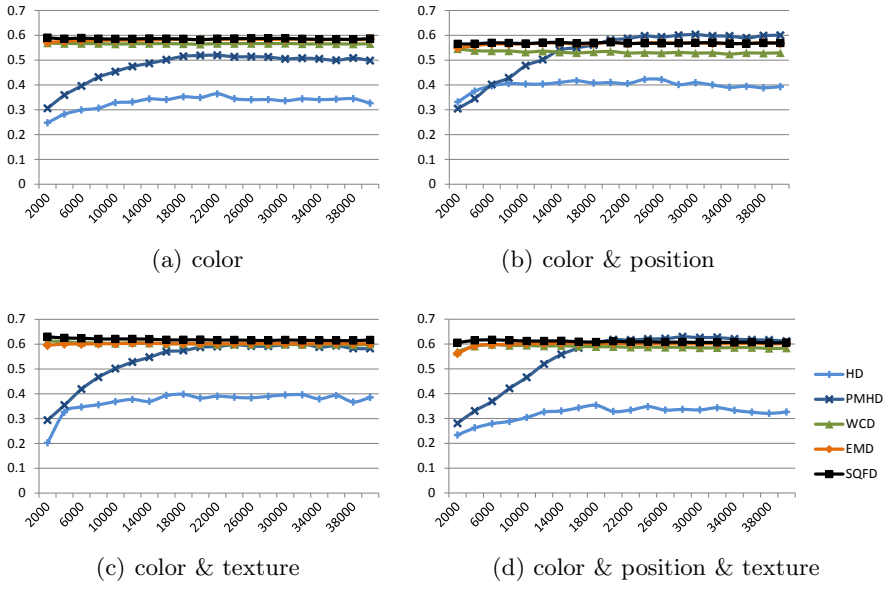
where $\mathbb{M}$ denotes the distribution of Mean Average Precision values with respect to the query modifying transformations $\Phi = \{\phi_1, \ldots, \phi_m\}$ applied to each query contained in the set of queries Q, i.e. $\mathbb{M} = \bigcup_{i=1}^{m} \{\mathrm{MAP}(\{\phi_i(q_1), \ldots, \phi_i(q_l)\},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

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 similarity measure's stability 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* generalized the Mean Average Precision measure by including the variance 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, as 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 adaptive similarity measures for the purpose of content-based image retrieval in the following section.

5 Experimental Evaluation

We evaluated the similarity measures' stability on the following benchmark image databases: the *Corel Wang* database [26] comprises 1,000 images which are classified into ten themes. The themes cover a multitude of topics, such as beaches, flowers, buses, food, etc. The *Coil 100* database [17] consists of 7,200 images classified into 100 different classes. Each class depicts one object photographed from 72 different directions. The *MIR Flickr* database [8] contains



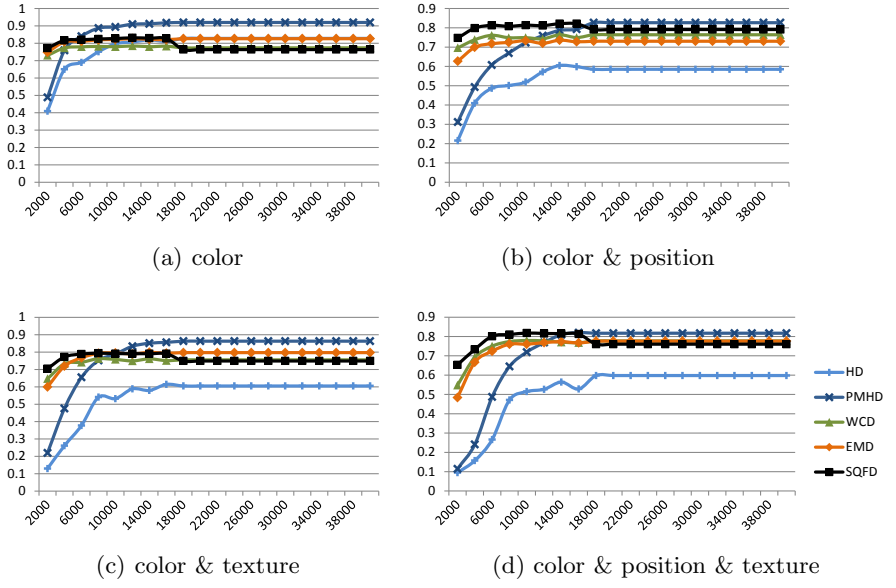
(a) color (b) color & position

(c) color & texture (d) color & position & texture

Fig. 2. Mean Average Precision values for the *Corel Wang* database based on different features: (a) color, (b) color & position, (c) color & texture, and (d) color & position & texture. The number of image pixels is varied between 2,000 and 40,000 pixels.

25,000 images downloaded from `http://flickr.com/` including textual annotations. The *101 objects* database [5] contains 9,196 images which are classified into 101 categories. Finally, we include the *ALOI* database [6] which is similar to the *Coil 100* database but comprises 72,000 images. The themes, classes, textual annotations, and categories are used as ground truth to measure precision and recall values [15] after each retrieved image. For the *MIR Flickr* database, we define virtual classes which contain all images sharing at least two common textual annotations and are used as ground truth.

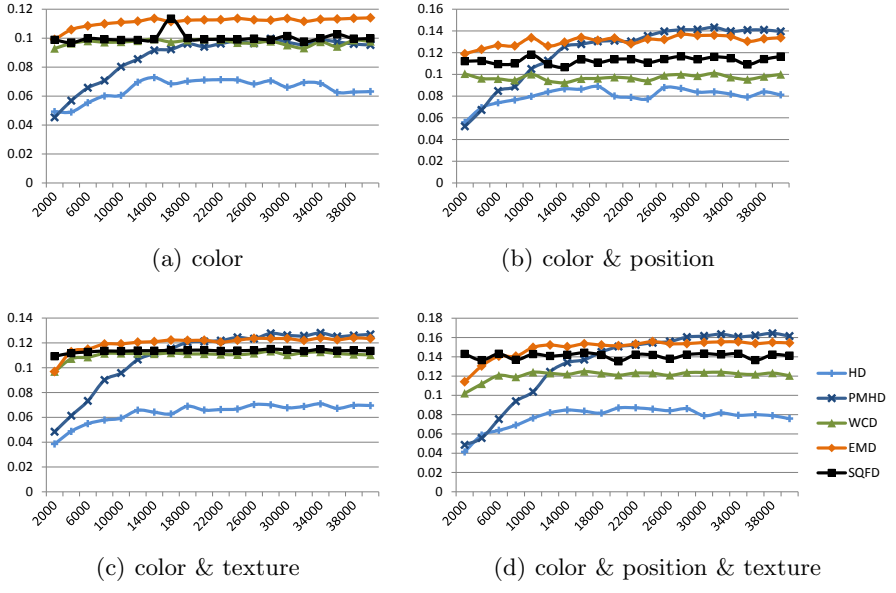
The resulting Mean Average Precision values, which are aggregated over 100 randomly selected queries for each combination of image database and similarity measure, are shown in Figures 2 to 6 where the number of image pixels considered for the extraction of color, position, and texture features is varied between 2,000 and 40,000. Thus the resulting query feature signatures generated by the adaptive k-means clustering algorithm vary in size between 1 and 115 centroids. The image databases always contain the feature signatures based on the clustering of 20,000 image pixels. In this way, the query modifying transformations are given by the change in cardinality of the query feature signatures, which is the most natural modification regardless of any specific local features. It can be seen in the figures, that the depicted mean average precision values depend on the applied feature spaces of the corresponding image database. In general, it turns out that the Hausdorff Distance (HD) and the Perceptually Modified Hausdorff Distance (PMHD) are very sensitive to change in feature signature quality on



(a) color (b) color & position

(c) color & texture (d) color & position & texture

Fig. 3. Mean Average Precision values for the *Coil 100* database based on different features: (a) color, (b) color & position, (c) color & texture, and (d) color & position & texture. The number of image pixels is varied between 2,000 and 40,000 pixels.



(a) color (b) color & position

(c) color & texture (d) color & position & texture

Fig. 4. Mean Average Precision values for the *101objects* database based on different features: (a) color, (b) color & position, (c) color & texture, and (d) color & position & texture. The number of image pixels is varied between 2,000 and 40,000 pixels.

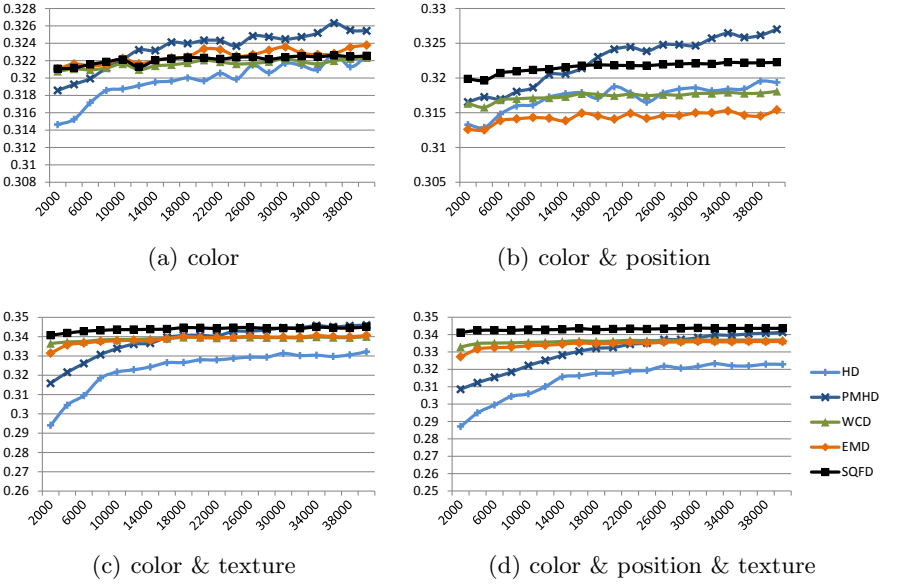


Fig. 5. Mean Average Precision values for the *MIR Flickr* database based on different features: (a) color, (b) color & position, (c) color & texture, and (d) color & position & texture. The number of image pixels is varied between 2,000 and 40,000 pixels.

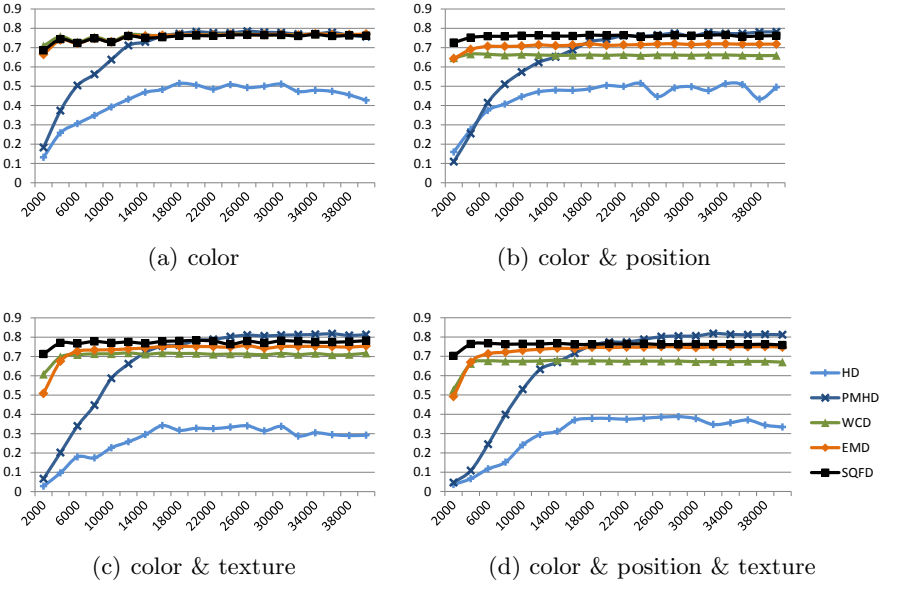


Fig. 6. Mean Average Precision values for the *ALOI* database based on different features: (a) color, (b) color & position, (c) color & texture, and (d) color & position & texture. The number of image pixels is varied between 2,000 and 40,000 pixels.

Table 1. Average Precision Stability (APS) regarding different features and sizes of the query feature signatures

database	$\mathbb{F}$	HD	PMHD	WCD	EMD	SQFD
Corel Wang	c	0.322	0.450	0.566	0.580	**0.585**
	c,p	0.392	0.488	0.530	0.563	**0.567**
	c,t	0.354	0.494	0.599	0.601	**0.616**
	c,p,t	0.310	0.493	0.584	0.594	**0.607**
Coil 100	c	0.712	0.802	0.765	**0.805**	0.763
	c,p	0.501	0.662	0.744	0.706	**0.784**
	c,t	0.481	0.678	0.730	**0.746**	0.743
	c,p,t	0.515	0.721	0.759	0.751	**0.769**
101 objects	c	0.065	0.086	0.097	**0.111**	0.100
	c,p	0.080	0.118	0.097	**0.130**	0.112
	c,t	0.063	0.107	0.110	**0.119**	0.113
	c,p,t	0.076	0.128	0.120	**0.147**	0.141
MIR Flickr	c	0.319	**0.323**	0.321	0.322	0.322
	c,p	0.317	**0.321**	0.317	0.314	**0.321**
	c,t	0.321	0.335	0.339	0.338	**0.344**
	c,p,t	0.311	0.327	0.336	0.334	**0.343**
ALOI	c	0.393	0.591	**0.747**	0.736	0.738
	c,p	0.411	0.546	0.658	0.698	**0.753**
	c,t	0.247	0.547	0.691	0.693	**0.761**
	c,p,t	0.269	0.517	0.645	0.686	**0.750**
average APS		0.323	0.437	0.488	0.499	**0.512**

the query side, while the Weighted Correlation Distance (WCD), Earth Mover's Distance (EMD), and Signature Quadratic Form Distance (SQFD) show more stable Mean Average Precision values. (A definition of these distance-based similarity measures can be found, for instance, in the work of Beecks et al. [2].)

In order to verify the observations mentioned above, we measured the Average Precision Stability: the results are reported in Table 1 where we highlighted the highest Average Precision Stability values of each row. On average, the Signature Quadratic Form Distance (SQFD) shows the highest Average Precision Stability values followed by the Earth Mover's Distance (EMD) and the Weighted Correlation Distance (WCD). In accordance with Figures 2 to 6, the Hausdorff Distance (HD) and the Perceptually Modified Hausdorff Distance (PMHD) show the lowest Average Precision Stability values, as they are more sensitive to query modifying transformations changing the query feature signatures' cardinalities.

To sum up, the experimental evaluation shows that the stability of the aforementioned similarity measures depends on the quality of the feature signatures appearing on the query side. While complex similarity models, such as the Earth Mover's Distance, Weighted Correlation Distance, and Signature Quadratic Form Distance, which take into account the complete structure of the feature signatures for the similarity value computation, are more robust against varying query signatures, the matching-based Hausdorff Distances suffer from query signatures deviating from the database signatures with respect to the cardinality. Thus the latter

are not feasible when the image models appearing on the query side significantly differ in size compared to those stored in the image database. In this case, the Earth Mover's Distance, the Weighted Correlation Distance, and particularly the Signature Quadratic Form Distance should be favored in order to obtain the highest stability.

6 Conclusions

We investigated the stability of the major adaptive similarity measures with respect to query modifying transformations. For this purpose, we defined the *Average Precision Stability* and evaluated the similarity measures' stability regarding the fundamental modification of size of the query feature signatures. As a result, the Signature Quadratic Form Distance shows the highest stability.

As future work, we plan to examine the Average Precision Stability of the similarity measures with respect to the qualities of photometric and geometric transformations.

References

1. Beecks, C., Uysal, M.S., Seidl, T.: Signature quadratic form distances for content-based similarity. In: Proc. ACM International Conference on Multimedia, pp. 697–700 (2009)
2. Beecks, C., Uysal, M.S., Seidl, T.: A comparative study of similarity measures for content-based multimedia retrieval. In: Proc. IEEE International Conference on Multimedia & Expo, pp. 1552–1557 (2010)
3. Beecks, C., Uysal, M.S., Seidl, T.: Signature quadratic form distance. In: Proc. ACM International Conference on Image and Video Retrieval, pp. 438–445 (2010)
4. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40(2), 1–60 (2008)
5. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples an incremental bayesian approach tested on 101 object categories. In: Proc. of the Workshop on Generative-Model Based Vision (2004)
6. Geusebroek, J.-M., Burghouts, G.J., Smeulders, A.W.M.: The Amsterdam Library of Object Images. International Journal of Computer Vision 61(1), 103–112 (2005)
7. Hu, R., Rüger, S., Song, D., Liu, H., Huang, Z.: Dissimilarity measures for content-based image retrieval. In: Proc. IEEE International Conference on Multimedia & Expo, pp. 1365–1368 (2008)
8. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: Proc. of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 39–43 (2008)
9. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.A.: Comparing Images Using the Hausdorff Distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(9), 850–863 (1993)
10. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20, 422–446 (2002)
11. Kent, A., Berry, M.M., Luehrs, F.U., Perry, J.W.: Machine literature searching viii. operational criteria for designing information retrieval systems. American Documentation 6(2), 93–101 (1955)

12. Leow, W.K., Li, R.: The analysis and applications of adaptive-binning color histograms. Computer Vision and Image Understanding 94(1-3), 67–91 (2004)
13. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications 2(1), 1–19 (2006)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
15. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
16. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
17. Nene, S., Nayar, S.K., Murase, H.: Columbia Object Image Library (COIL-100). Technical report, Department of Computer Science, Columbia University (1996)
18. Park, B.G., Lee, K.M., Lee, S.U.: Color-based image retrieval using perceptually modified Hausdorff distance. Journal on Image and Video Processing 2008, 1–10 (2008)
19. Rubner, Y., Puzicha, J., Tomasi, C., Buhmann, J.M.: Empirical evaluation of dissimilarity measures for color and texture. Computer Vision and Image Understanding 84(1), 25–43 (2001)
20. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)
21. Sebe, N., Lew, M.S., Zhou, X., Huang, T.S., Bakker, E.M.: The state of the art in image and video retrieval. In: Proc. ACM International Conference on Image and Video Retrieval, pp. 1–8 (2003)
22. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, pp. 1470–1477 (2003)
23. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
24. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: a survey. Foundations and Trends in Computer Graphics and Vision 3(3), 177–280 (2008)
25. van Rijsbergen, C.J.: Information Retrieval. Butterworth (1979)
26. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(9), 947–963 (2001)