Visual Image Search: Feature Signatures or/and Global Descriptors

Jakub Lokoč[1], David Novák[2], Michal Batko[2], and Tomáš Skopal[1]

[1] SIRET Research Group, Faculty of Mathematics and Physics,
Charles University in Prague
{lokoc,skopal}@ksi.mff.cuni.cz
[2] Faculty of Informatics, Masaryk University in Brno
{david.novak,batko}@fi.muni.cz

Abstract. The success of content-based retrieval systems stands or falls with the quality of the utilized similarity model. In the case of having no additional keywords or annotations provided with the multimedia data, the hard task is to guarantee the highest possible retrieval precision using only content-based retrieval techniques. In this paper we push the visual image search a step further by testing effective combination of two orthogonal approaches – the MPEG-7 global visual descriptors and the feature signatures equipped by the Signature Quadratic Form Distance. We investigate various ways of descriptor combinations and evaluate the overall effectiveness of the search on three different image collections. Moreover, we introduce a new image collection, TWIC, designed as a larger realistic image collection providing ground truth. In all the experiments, the combination of descriptors proved its superior performance on all tested collections. Furthermore, we propose a re-ranking variant guaranteeing efficient yet effective image retrieval.

1 Introduction

With the increasing volumes of multimedia data available over the internet, the Content-based Image Retrieval Systems (CBIR) [10,11] steadily become more and more important. Even though for some of the data an annotation is available, the content-based paradigm (possibly combined with the keyword search) might provide more precise retrieval than the keyword search alone. This fact was recently confirmed by Google that added content-based image search to the classic keyword image search engine. However, unlike keyword search, the technology of content-based image retrieval is extremely diverse as there have been thousands models proposed and even more studies performed [10]. The differences reflect various demands, such as general image-matching technologies and applications vs. very specialized systems tuned for a specific application, use various feature extraction types designed for measuring global or local similarity, and so on. The image descriptors span from image fingerprints (hashes) for near-duplicate search, over local image features, to descriptors of global features.

G. Navarro and V. Pestov (Eds.): SISAP 2012, LNCS 7404, pp. 177–191, 2012.

In this paper, we deal with MPEG-7 global visual descriptors and so-called image feature signatures. The MPEG-7 visual descriptors use standardized description of image content that proved to provide good retrieval effectiveness in image retrieval applications. Their main property is they describe global image features, such as color, texture or shape distribution, among others. On the other hand, the recently proposed feature signatures allow to aggregate local features into a compact form. It has been shown that feature signatures provide more flexible similarity search than MPEG-7 descriptors, while they offer less complex matching than local features developed for image classification, e.g., SIFTs.

Paper Contribution
Both MPEG-7 descriptors and feature signatures have their strong and weak points. In this paper we compare and synergistically combine MPEG-7 descriptors with image feature signatures in order to reach ultimate effectiveness. In particular, we

- evaluate on three different image collections the effectiveness of standard global visual descriptors (and their combinations) and complex feature signatures used with the signature quadratic form distance (SQFD),
- combine the MPEG-7 global descriptors with the feature signatures in various ways which improves the overall effectiveness of the search on all tested collections,
- employ the re-ranking concept, such that the MPEG-7 descriptors are used for "cheap" pre-selection of image candidates and then the (rather small) result is re-ranked using the time-consuming SQFD based on feature signatures, resulting thus in a very efficient search mechanism,
- and introduce a new image collection for CBIR effectiveness evaluation.

2 Preliminaries and Related Work

When searching multimedia databases in a content-based way, users issue similarity queries by selecting multimedia objects or by sketching the intended object contents. Given an example multimedia object or sketch q, the multimedia database $\mathbb{S} \subset \mathbb{U}$ (where $\mathbb{U}$ is the object universe) is searched for the most related objects with respect to the query by measuring the similarity between the query and each database object by means of a distance function δ. As a result, the multimedia objects with the lowest distance to the query are returned to the user. In particular, a *range query* (q, r), $q \in \mathbb{U}$, $r \in \mathbb{R}^+$, reports all objects in $\mathbb{S}$ that are within a distance r to q, that is, $(q, r) = \{x \in \mathbb{S} \mid \delta(x, q) \leq r\}$. The subspace defined by q and r is called the *query ball*. Another popular similarity query is the k *nearest neighbors query* (k-NN(q)). It reports the k objects from $\mathbb{S}$ closest to q. That is, it returns the set $\mathbb{C} \subseteq \mathbb{S}$ such that $|\mathbb{C}| = k$ and $\forall x \in \mathbb{C}, y \in \mathbb{S} - \mathbb{C}, \delta(x, q) \leq \delta(y, q)$. The k-NN query also defines a query ball (q, r), but the distance r to the k^{th} NN is not known beforehand. In the following paragraphs, we describe two different model representations used in this paper.

2.1 MPEG-7 Global Visual Descriptors

The global visual descriptors are the fundamental instruments to measure the overall similarity of the digital images' content. In this work, we use five well-established descriptors from the MPEG-7 standard [24] that capture various image characteristics. There is a function defined for each of the descriptors [21] to measure the *distance* (dissimilarity) δ between two instances of that descriptor.

Scalable Color is derived from a color histogram in the Hue-Saturation-Value color space with fixed space quantization. We used the 64 coefficients version of this descriptor. The distance between two scalable color instances is measured by the L_1 metric (sum of absolute differences).

Color Structure aims at identifying localized color distributions using a 8×8 pixels structuring matrix that slides over the image. This descriptor can distinguish between two images having similar amount of pixels of a specific color, if structures of these pixels differ in these images. The L_1 metric is used to compute descriptors distances.

Color Layout descriptor is obtained by applying the Discrete cosine transform on a 2-D array (usually 8×8 blocks) of local representative colors in three color channels (Y, Cb, and Cr). The distance between two objects is computed as a sum of L_2 distances in each of the three color space components.

Edge Histogram represents the local-edge distribution in the image. The image is subdivided into 4×4 sub-images and edges in each sub-image are categorized into five types: vertical, horizontal, 45° diagonal, 135° diagonal, and non-directional edges. This results in 80 coefficients (5 values for each of the 16 sub-images) representing the local edge histograms. Further, the semi-global and the global histograms can be computed based on the local histogram and the distance is computed as a sum of weighted sub-sums of absolute differences for the local, semi-global and global histograms.

Region Shape descriptor considers the whole region of the shapes on the image. The descriptor works by "decomposing" the shape into a number of orthogonal 2-D basis functions defined by the Angular Radial Transformation (ART) [24]. The descriptor is a vector of normalized magnitudes of the ART coefficients and the distance is calculated using the L_1 norm.

2.2 Descriptors Consisting of Local Features

The conventional feature descriptors, such as MPEG-7 visual descriptors, aggregate and store these properties in *feature histograms*, which can be compared by vectorial distances [17,26]. The problem is, that for both simple and complex images there is the same number of bins, which does not reflect the complexity of the images. From this point of view, the *feature signatures* are more flexible choice to describe the image content.

Feature Signatures. Unlike conventional feature histograms, feature signatures are frequently obtained by clustering the objects' properties, such as color, position, texture, or other more complex features [12,23], within some feature space and storing the cluster representatives and weights. Thus, given a feature space $\mathbb{F}$, the *feature signature* S^o of a multimedia object o is defined as a set of tuples from $\mathbb{F} \times \mathbb{R}^+$ consisting of representatives $r^o \in \mathbb{F}$ and weights $w^o \in \mathbb{R}^+$.

Fig. 1. Three example images with their corresponding feature signature visualizations

We depict an example of image feature signatures according to a feature space comprising position, color and texture information, i.e. $\mathbb{F} \subseteq \mathbb{R}^7$, in Figure 1. For this purpose, we applied a k-means clustering algorithm where each representative $r_i^o \in \mathbb{F}$ corresponds to the centroid of the cluster $\mathcal{C}_i^o \subseteq \mathbb{F}$, i.e., $r_i^o = \frac{\sum_{f \in \mathcal{C}_i^o} f}{|\mathcal{C}_i^o|}$, with relative frequency $w_i^o = \frac{|\mathcal{C}_i^o|}{\sum_i |\mathcal{C}_i^o|}$. We depict the feature signatures' representatives by circles in the corresponding color. The weights are reflected by the diameter of the circles. As can be seen in this example, feature signatures adjust to individual image contents by aggregating the features according to their appearance in the underlying feature space.

Signature Quadratic Form Distance. The Signature Quadratic Form Distance (SQFD) [6] is an adaptive distance-based similarity measure, generalizing the classic vectorial Quadratic Form Distance (QFD) [16] for feature signatures. It is defined as follows.

Definition 1 (SQFD). *Given two feature signatures* $S^q = \{\langle r_i^q, w_i^q \rangle\}_{i=1}^n$ *and* $S^o = \{\langle r_i^o, w_i^o \rangle\}_{i=1}^m$ *and a similarity function* $f_s : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ *over a feature space* $\mathbb{F}$*, the signature quadratic form distance* SQFD_{f_s} *between* S^q *and* S^o *is defined as:*

$$\mathrm{SQFD}_{f_s}(S^q, S^o) = \sqrt{(w_q \mid -w_o) \cdot A_{f_s} \cdot (w_q \mid -w_o)^T},$$

where $A_{f_s} \in \mathbb{R}^{(n+m)\times(n+m)}$ *is the similarity matrix arising from applying the similarity function* f_s *to the corresponding feature representatives, i.e.,* $a_{ij} = f_s(r_i, r_j)$*. Furthermore,* $w_q = (w_1^q, \ldots, w_n^q)$ *and* $w_o = (w_1^o, \ldots, w_m^o)$ *form weight vectors, and* $(w_q \mid -w_o) = (w_1^q, \ldots, w_n^q, -w_1^o, \ldots, -w_m^o)$ *denotes the concatenation of weights* w_q *and* $-w_o$*.*

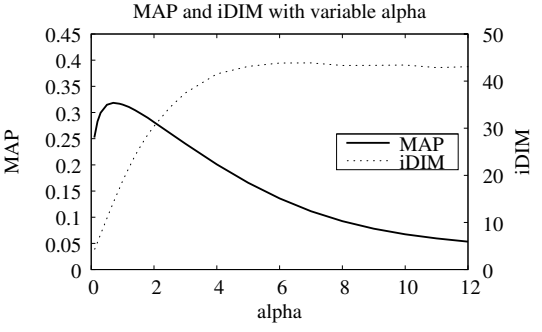

Fig. 2. The impact of α on the mean average precision (MAP) and intrinsic dimensionality (iDIM)

The similarity function f_s is used to determine similarity values between all pairs of representatives from the feature signatures. In our implementation, we use the similarity function $f_s(r_i, r_j) = e^{-\alpha L_2(r_i, r_j)^2}$, where α is a constant for controlling the precision-indexability tradeoff, as investigated in previous works [4,19], and L_2 denotes the Euclidean distance. In particular, the lower values of α lead to better indexability (allowing fast search), that is, to lower values of so-called *intrinsic dimensionality* (iDIM) [8]. However, with lower values of the α parameter also the *mean average precision* (MAP) decreases, see an example for TWIC database in Figure 2. On the contrary, the best mean average precision values can be reached for already high α (e.g., $\alpha > 0.5$ in the figure), where the SQFD space is no longer indexable. In such cases the parallel implementation could be the only feasible way to significantly speedup the search, especially when GPU processing is employed [18]. In the following section we briefly summarize the indexing methods used for efficient retrieval.

2.3 Efficiency of the Retrieval

In this section, we briefly analyze how demanding is indexing and searching of the above mentioned descriptors, while we mention ways of how to speed up the search. Details can be found in the referenced literature.

MPEG-7 Visual Descriptors. The described MPEG-7 visual descriptors are a standard means for measuring global visual similarity of images. Indexing and

searching of such data can be relatively efficient because the representations of the features is not very space demanding (all five mentioned descriptors together require about 1 kB of memory) and all the respective distance functions are based on L_p metrics. Individual descriptors can be combined by a (weighted) sum of their respective distances and the result remains a metric space. On average, the time of a single distance computation of this descriptor combination is about 0.01 ms.

Number of recent works on metric-based indexing and searching presented efficiency experiments on combination of several global features [2,14,25]. The results indicate that evaluation of k-NN on such spaces is relatively efficient for both precise and approximate similarity search: over 50 % of the indexed data can be pruned by sophisticated precise metric access methods and about 90 % of the precise k-NN answer can be obtained by accessing 5–10 % of the indexed data by sophisticated approximate techniques.

Indexing Feature Signatures with SQFD. When processing content-based similarity queries by the naïve sequential search, the SQFD distance has to be evaluated for each database object individually. Unlike the cheap L_p distances, the SQFD is of more than quadratic time complexity, so the sequential search, sometimes acceptable for L_p distances, is impractical for SQFD even on a moderately sized database. Although it has been shown that the SQFD is a generalization [7] of the well-known Quadratic Form Distance [16], recent approaches indexing the data by a homeomorphic mapping into the Euclidean space [27] cannot be applied to the SQFD, as the similarity matrix changes from computation to computation.

Nevertheless, recent papers showed that SQFD can be indexed by metric access methods [4] and ptolemaic indexing [19], achieving a speed-up of up to two orders of magnitude with respect to the sequential scan.

2.4 Descriptor Effectiveness: Related Work

There are a number of works studying the effectiveness of global MPEG-7 descriptors and their combination for both general visual image search [13,29,3,1] and more specific application [9,28]. We can draw the following general conclusion from these studies: the descriptors can well serve for a relatively fast global visual image search and they were successfully applied in a number of applications. Specific selection or combination of descriptors depends on specific characteristics of the dataset.

Visual feature signatures, especially in combination with the SQFD measure, draw attention recently and effectiveness of this approach was tested on several CBIR collections [7,5]. These works focus on comparison of various similarity measures for this type of descriptors and they identified SQFD as superior. To the best of our knowledge, there exists no work that would compare effectiveness of feature signatures (with SQFD) and standard MPEG-7 descriptors for global visual search or that would combine these two approaches.

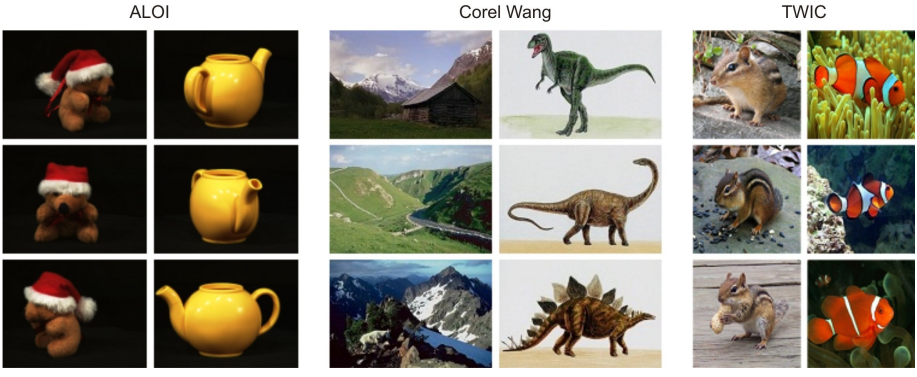

Fig. 3. The datasets used in the experiments, each dataset represented by two classes/topics

3 Global Descriptors and Feature Signatures

The following section is the key part of the paper. First, we describe the experiment settings, i.e., datasets, used descriptors and employed evaluation metrics. Then we evaluate the effectiveness of the individual approaches, plus we evaluate and describe various combinations of global descriptors and feature signatures.

3.1 Experiment Settings

Datasets. To conduct the experiments, first the ALOI dataset [15] comprising 72,000 images and the Corel Wang dataset [30] comprising 1,000 images were considered. Both datasets provide the ground truth in the form of classes containing particular images. The ALOI dataset consists of 1,000 classes where each class represents one object captured under various viewing angles. Six example images representing two classes in the ALOI dataset are depicted in the first column of Figure 3. Since all the images have black background which reduces the noise information and the classes are very homogeneous, the similarity search task in ALOI dataset is quite simple. The Corel Wang dataset consists of more heterogeneous images selected from ten different topics (see the second column of Figure 3). Such dataset can verify the proposed methods more thoroughly.

However, the Corel Wang dataset is quite small and not all images in particular topics are visually similar. Therefore, we have decided to create and introduce a new dataset called *Thematic Web Images Collection* (TWIC) comprising 11,555 images divided among 200 classes [20]. The TWIC dataset is intended as an alternative to ALOI – each class consists of visually similar objects but the background is heterogeneous. Six images representing two classes in the TWIC dataset are depicted in the third column of Figure 3. We may observe that in one class there is one central object on various backgrounds, which more corresponds to real requirements of the visual similarity search tasks.

To create the TWIC dataset, we have first selected several domains (e.g., Buildings, Flags, Mammals, Ocean, etc.) and for each domain we have selected around fifty keywords from that domain. Having these several hundreds of keywords where each keyword represents one image class, we have started to query the google images engine. Such keywords that created a homogeneous google image search result[1] were saved, i.e., the keyword and first two-hundred links to the corresponding images. Then, we have manually filtered all images that were not visually coherent from each image class and selected only the classes containing more than fifty objects obtaining finally 11,555 images divided among 200 classes (keywords). For more details see [20].

Descriptors. In the experiments, we used the five MPEG-7 descriptors described in Section 2.1 [24] together with the recommended distance measures. Standard XM library was used for extraction [21].

To create feature signatures, we have extracted seven-dimensional features $(L, a, b, x, y, \chi, \epsilon) \in \mathbb{F}$ including color (L, a, b), position (x, y), contrast χ, and entropy ϵ information (as suggested in [5]). We obtained one feature signature for every single image, where the signatures varied in size between 12 and 48 feature representatives. On average, a feature signature consisted of 30 representatives (i.e., 240 numbers per signature).

We combine individual descriptors by a (weighted) sum of their respective distances. As the individual descriptors (including the feature signatures) with the described distance functions form metric spaces, the combined space is also metric, which is important for efficient indexing. Individual distance components can be normalized and weighted as we will see further in Section 3.2.

Querying and Evaluation Metrics. Effectiveness of individual descriptors and their combinations is evaluated by precision of query answers. An image in the answer is considered part of precise answer, if it belongs to the same image class as the query image. Namely, we executed: 1,000 queries for ALOI dataset (each query from one image class), 100 queries for Corel Wang dataset (ten from each of the ten topics) and 200 queries for TWIC dataset (each from one image class). Within each dataset, we calculated *mean average precision* (MAP) [22] over the set of queries and also *average precision* for k-NN results with variable k. Further, we measured *intrinsic dimensionality* (iDIM) from the distance distribution of respective descriptor space [8].

3.2 Effectiveness of Individual Approaches

In this section, we describe the effectiveness of individual descriptors on all three datasets under test. Table 1 summarizes the results for all five global MPEG-7 descriptors individually and for the feature signatures with the SQFD distance.

As expected, the overall values of MAP differ significantly for individual datasets: ALOI is relatively uncomplicated dataset with MAP reaching values over 0.7 even

[1] The result contained many visually similar images.

Table 1. Mean Average Precision (MAP) of individual descriptors and their intrinsic dimensionality (iDIM)

Individual descriptors	ALOI		Corel Wang		TWIC	
	MAP	iDIM	MAP	iDIM	MAP	iDIM
Color Layout	0.37	1.3	0.43	4.2	0.17	4
Color Structure	0.78	2	0.5	7.7	0.20	6
Edge Histogram	0.28	2	0.4	6.8	0.15	5
Region Shape	0.15	0.7	0.25	2.4	0.07	2.5
Scalable Color	0.70	2.7	0.48	9.2	0.15	7
SQFD	0.73	3.7	0.49	15	0.32	18.5
MPEG7 combination	0.78	3	0.57	13	0.30	13.7

for single descriptors; we can observe MAP up to 0.5 for Corel Wang; and TWIC (as the most realistic dataset) has MAP values up to 0.2 for single global descriptors and 0.32 for feature signatures with SQFD. The last row of the table shows results for five combined MPEG-7 descriptors, each normalized by its maximum distance and summarized. As expected, this measure outperforms individual descriptors and, for ALOI and Corel Wang, it is better than SQFD.

The second columns for each dataset depict intrinsic dimensionality for the respective descriptor spaces. We can see that ALOI descriptors have significantly smaller iDIM which is caused by small actual visual difference between images in the dataset (all images depict single object isolated on a black background). Also, individual iDIM values often correspond with the effectiveness (MAP) of respective descriptors. As the feature signatures with SQFD cover several low-level features, the iDIM of this space is by far the highest.

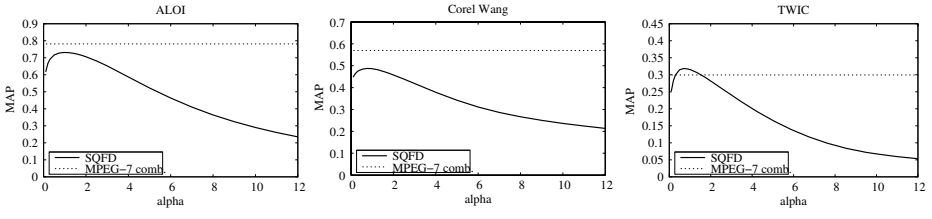

Fig. 4. MAP using global descriptor combination and SQFD with variable α

As discussed in Section 2.2, SQFD allows to control the precision-indexability tradeoff by the parameter α used in the similarity function $f_s(r_i, r_j) = e^{-\alpha L_2(r_i, r_j)^2}$. In Figure 4, we can observe the MAP results for varying α in all three datasets. The optimal value of all of them is around value 1, which is caused by the fact that the feature extraction method was used with the same parameters. Too small or too big α results in less diverse values in the similarity matrix[2] and

[2] In the limit case, the resulting similarity matrix can be either nearly diagonal or unitary.

thus in the loss of information useful for similarity. Setting $\alpha = 1$ was used for all other experiments in this section (also in Table 1). For comparison, the figure also shows MAP of the combination of global MPEG-7 descriptors mentioned above.

We can summarize these results as follows: Effectiveness of the feature signatures with SQFD is mostly better than a single MPEG-7 global descriptor and comparable with the MPEG-7 descriptor combination. The question remains, whether better results can be reached by fusion of both approaches.

Table 2. Mean Average Precision (MAP) of various descriptor combinations: the distance measure of the combination is either (1) pure sum of its component sub-distances, or (2) sum of these sub-distances weighted by the components iDIM

Effectiveness (MAP)						
	ALOI		**Corel Wang**		**TWIC**	
Combination of descriptors	**sum**	**iDIM**	**sum**	**iDIM**	**sum**	**iDIM**
MPEG7 combination	0.78	0.80	0.57	0.58	0.30	0.30
Color Layout + SQFD	*0.71*	0.74	0.50	0.51	0.33	0.34
Color Structure + SQFD	0.82	0.80	0.56	0.55	0.34	0.34
Edge Histogram + SQFD	*0.71*	0.74	0.52	0.52	0.35	0.34
Region Shape + SQFD	*0.70*	0.73	*0.45*	0.49	*0.29*	0.32
Scalable Color + SQFD	0.83	0.83	0.55	0.54	0.33	0.34
MPEG7 combination + SQFD	0.81	**0.83**	0.58	**0.59**	0.37	**0.38**

3.3 Approach Combinations

The previous section showed that combination of MPEG-7 descriptors and feature signatures with SQFD exhibit relatively similar effectiveness and their combination might be advantageous. All the descriptor spaces in question are metric and it is important for indexing and searching that the combination space preserve the metric properties. Therefore, we again decided to combine the spaces by a (weighted) sum of their respective distances. The first row of Table 2 shows again results of the MPEG-7 descriptor combination and then combinations with SQFD are presented.

For each dataset, the first column always means the pure sum of respective descriptors (normalized by maximum distances in each descriptor space), i.e.

$$\delta_{D_1+\cdots+D_n}(X,Y) = \sum_{i=1}^{n} \frac{\delta_{D_i}(X_{D_i}, Y_{D_i})}{\text{max\_dist}_{D_i}}, \tag{1}$$

where $D_1, \ldots, D_n$ denote individual descriptors (e.g. Color Layout or feature signatures with SQFD), X_{D_i} and Y_{D_i} denote values of the D_i descriptor of images X and Y, respectively, and δ_{D_i} is the distance function used with descriptor D_i (see Sections 2.1 and 2.2). When we compare MAP values from Tables 1 and 2, practically any combination (Table 2) reached higher MAP than its components

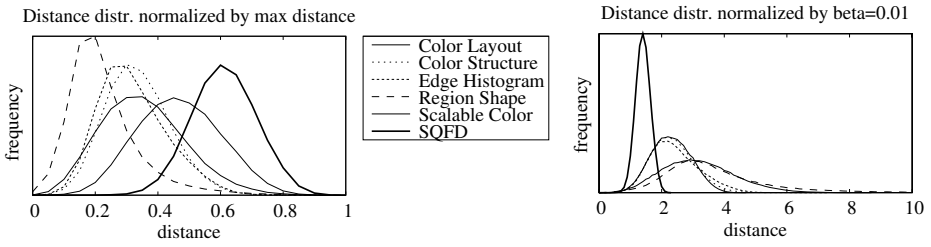

Fig. 5. Distance distribution of individual descriptors from TWIC: (a) normalized by maximum distance, (b) normalized by $\beta = 0.01$

(Table 1); the exceptional combinations (that worsened the MAP) are emphasized by italic numbers in Table 2. These exceptions always involve descriptors with extremely poor MAP (from Table 1).

A recent work [1] addressed the question of weight selection for combinations of metric visual descriptors. The authors studied the distance distribution of descriptor components normalized by the maximum distances to interval $[0, 1]$ – see distance histogram in Figure 5 (left). In general, descriptors with larger average distances would influence the combination sum more significantly, which might be a potential issue. We can try to overcome this by normalizing each descriptor D_i by a distance τ_{D_i} smaller than maximum max_dist_{D_i}. Looking at the distance histograms, the authors proposed to determine τ for each descriptor so that it corresponds to a certain fixed *percentage* β of the smallest distances in the histogram (see [1] for details). The effect of this normalization for $\beta = 1\,\%$ is depicted in Figure 5 (right) – the beginnings of individual curves are very close, which should improve the effectiveness of the combination 5. Following this idea [1], we repeated all experiments with this normalization using $\beta = 1\%$ and $\beta = 1‰$, but the results were always slightly *worse* than with normalization by maximum distance. We plan to investigate this area even deeper in the future.

Nevertheless, we successfully applied another weight-tuning technique to improve the overall effectiveness of descriptor combination. As we mentioned already, we can observe a correlation between effectiveness of individual descriptors and their intrinsic dimensionality iDIM (see Table 1). In general, the iDIM tries to quantify the complexity of the data space and the difficulty to index such dataset using a metric access method [8]. The observed iDIM-MAP correlation can be naturally explained so that more complex descriptors have higher iDIM and their search effectiveness is higher. Barrios et al. made similar observation and they determine individual descriptor weights so that the iDIM of the combined space is maximized (finding a local maxima) [1]. We propose an alternative approach that builds directly on the observed correlation – we weight individual descriptors in the combination (1) by the respective iDIM:

$$\delta^{\text{iDIM}}_{D_1+\cdots+D_n}(X, Y) = \sum_{i=1}^{n} \text{iDIM}_{D_i} \cdot \frac{\delta_{D_i}(X_{D_i}, Y_{D_i})}{\text{max\_dist}_{D_i}}. \tag{2}$$

The MAP for these experiments on all datasets are in the second columns in Table 2 (denoted as *iDIM weighted*) and we can see that practically all the MAP values improved – especially the overall maxima achieved by combination of all global descriptors and feature signatures with SQFD (last row).

Comparing these values of MAP (printed in bold) with values of the MPEG-7 combination and values of pure signatures with SQFD, this best combination resulted in improvement for all three datasets from 0.02 points (Corel Wang) to 0.08 points (TWIC). We consider these results as a success, because the MAP measure is always very difficult to improve, especially when the ground truth for each query forms only a small fraction of the whole collection (it is 0.5 % for the TWIC dataset, on average).

Table 3. Mean Average Precision (MAP) of searching k-NN using MPEG-7 desc. combination and then re-ranking of the kNN results using combination of MPEG-7 descriptors and signatures with SQFD

Effectiveness (MAP) of SQFD re-ranking						
	ALOI		**Corel Wang**		**TWIC**	
Combination of descriptors	**sum**	**iDIM**	**sum**	**iDIM**	**sum**	**iDIM**
MPEG7 100NN, re-rank MPEG7+SQFD	0.79	0.81	0.44	0.45	0.28	0.28
MPEG7 200NN, re-rank MPEG7+SQFD	0.80	0.82	0.52	0.53	0.32	0.32
MPEG7 300NN, re-rank MPEG7+SQFD	0.81	0.82	0.55	0.56	0.33	0.33
MPEG7 500NN, re-rank MPEG7+SQFD	0.81	0.83	0.57	0.58	0.35	0.35
MPEG7 1000NN, re-rank MPEG7+SQFD	0.81	0.83	–	–	0.36	0.37

3.4 Re-Ranking by SQFD

As mentioned in Section 2.3, feature signatures with SQFD form a significantly more difficult data space for indexing and searching than MPEG-7 global descriptors accompanied with relatively cheap L_p-based distances. For larger datasets, it could be very time consuming to index the collection according to our most effective combination – MPEG-7 descriptors and feature signatures with SQFD.

This led us to the following schema: We index the data by the MPEG-7 descriptor combination, evaluate the k-NN(q) on such index, and re-rank these k images according to distance "MPEG-7 combination + SQFD" evaluated with respect to query image q. Results of this approach are summarized in Table 3 for $k = 100$, 200, 300, 500, and 1,000 – again for variants with individual weights equal to 1 and to respective iDIM values.

When we compare these results with the last row of Table 2 (MPEG-7 combination + SQFD), we can see that with growing k, the MAP gets very close to these maximal values. We can conclude, that even re-ranking by combination of MPEG-7 descriptors and feature signatures can improve the overall quality of the visual search almost as doing the whole search directly by the combination. However, the direct approach increases the search costs insignificantly.

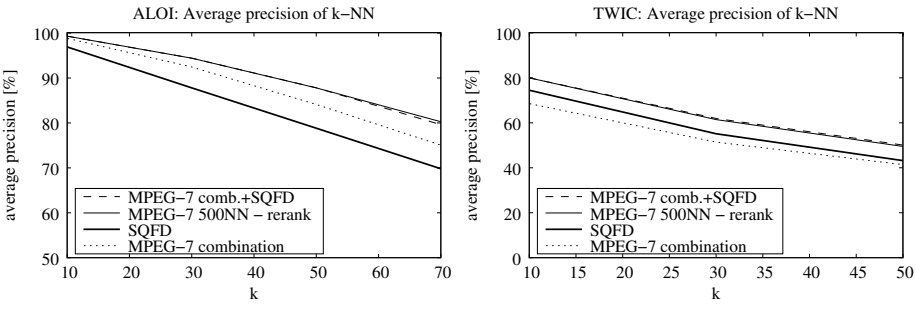

Fig. 6. ALOI and TWIC: Average precision of k-NN results for different approaches

The MAP is a very complex measure, but it does not necessarily give a good intuition of how good would be results of a real k-NN for a reasonable k; especially for larger datasets, improving MAP can be difficult even though precision of standard k-NN answers can be very high. Therefore, we measured also the k-NN *average precision* for variable k and for the most important search approaches introduced above. See Figure 6 for these results on ALOI and TWIC datasets. Always, we present results for the combination of MPEG-7 descriptors, feature signatures with SQFD, the MPEG-7 combination + SQFD, and MPEG-7 500-NN re-ranked by MPEG-7 combination + SQFD. We can see that for ALOI, the precision is practically 100 % for 10-NN and falls down only to 80 % for 70-NN using the best approaches. Improvement of our combination in comparison with pure SQFD is up to 10 %. We can also see, that the average precision of the re-ranking approach is practically identical as the MPEG-7+SQFD approach. Note that the x-axis ends at 70 because that is the size of image classes for ALOI (analogously for other figures). For Corel Wang and TWIC dataset, the average precision is between 80 % and 50 % and the improvement of the combinations with respect to individual approaches is also about 10 %.

4 Conclusions and Future Work

In this paper, we combined feature signatures with MPEG-7 global visual descriptors to improve the performance of content-based image search engines. We proposed several techniques on how to combine these two orthogonal approaches and experimentally shown the positive synergistic effect of the combination. Hence, we could conclude that it is profitable to utilize both, MPEG-7 descriptors and feature signatures, because they complement each other and improve the quality of the retrieval. We also introduced a new (and more realistic) test image collection providing ground truth comprising images obtained via Google image search. The collection was designed to fill the gap in the realistic multimedia test collections with ground truth. Finally, we proposed a re-ranking variant

guaranteeing efficient yet effective image retrieval. In the future, we plan to investigate feature signature extraction that profits from the descriptor combination. Also, we plan to compare our approach with the Bag-of-Words method as another standard approach successfully applied in multimedia retrieval. Additional experiments on larger datasets might show that MPEG-7 performance deteriorate faster than the performance of feature signatures. Hence, our re-ranking approach could contribute to more robust behavior on large databases.

Another interesting theoretical problem to inspect is the SQFD behavior under varying alpha parameter. It seems that there is an optimal value of alpha parameter for the precision, while with lower alpha we get always lower intrinsic dimensionality (iDIM is monotonously dependent on alpha). The relation between alpha parameter and the intrinsic dimensionality should be theoretically explained to provide more clues for the design of the SQFD distance spaces.

Acknowledgments. This research has been supported in part by Czech Science Foundation (GACR) projects P202/11/0968, P202/12/P297, 103/10/0886 and P202/10/P220.

References

1. Barrios, J.M., Bustos, B.: Automatic weight selection for multi-metric distances. In: Proceedings of the Fourth International Conference on SImilarity Search and APplications, SISAP 2011, pp. 61–68. ACM Press, New York (2011)
2. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubsky, J., Zezula, P.: Building a web-scale image similarity search system. Multimedia Tools and Applications 47(3), 599–629 (2010)
3. Batko, M., Kohoutkova, P., Novak, D.: CoPhIR Image Collection under the Microscope. In: Second International Workshop on Similarity Search and Applications (SISAP 2009), pp. 47–54. IEEE (2009)
4. Beecks, C., Lokoč, J., Seidl, T., Skopal, T.: Indexing the signature quadratic form distance for efficient content-based multimedia retrieval. In: Proc. ACM Int. Conf. on Multimedia Retrieval, pp. 24:1–24:8 (2011)
5. Beecks, C., Uysal, M., Seidl, T.: A comparative study of similarity measures for content-based multimedia retrieval. In: 2010 IEEE International Conference on Multimedia and Expo (ICME), pp. 1552–1557. IEEE (2010)
6. Beecks, C., Uysal, M.S., Seidl, T.: Signature quadratic form distance. In: Proc. ACM CIVR, pp. 438–445 (2010)
7. 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)
8. Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J.L.: Searching in metric spaces. ACM Computing Surveys 33(3), 273–321 (2001)
9. Coimbra, M.T., Cunha, J.P.S.: MPEG-7 Visual Descriptors -Contributions for Automated Feature Extraction in Capsule Endoscopy. IEEE Transactions on Circuits and Systems for Video Technology 16(5), 628–637 (2006)
10. 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)
11. Deb, S.: Multimedia Systems and Content-Based Image Retrieval. Information Science Publ. (2004)

12. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. Information Retrieval 11(2), 77–107 (2008)
13. Eidenberger, H.: How good are the visual MPEG-7 features. In: SPIE Visual Communications and Image Processing Conference, vol. 5150, pp. 476–788. SPIE (2003)
14. Esuli, A.: PP-Index: Using permutation prefixes for efficient and scalable approximate similarity search. In: Proceedings of LSDS-IR 2009 (2009)
15. Geusebroek, J.-M., Burghouts, G.J., Smeulders, A.W.M.: The Amsterdam Library of Object Images. IJCV 61(1), 103–112 (2005)
16. Hafner, J., Sawhney, H.S., Equitz, W., Flickner, M., Niblack, W.: Efficient color histogram indexing for quadratic form distance functions. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 729–736 (1995)
17. 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)
18. Kruliš, M., Lokoč, J., Beecks, C., Skopal, T., Seidl, T.: Processing the signature quadratic form distance on many-core gpu architectures. In: Proceedings International Conference on Information and Knowledge Management, pp. 2373–2376 (2011)
19. Lokoč, J., Hetland, M., Skopal, T., Beecks, C.: Ptolemaic indexing of the signature quadratic form distance. In: Proceedings of the Fourth International Conference on SImilarity Search and Applications, pp. 9–16. ACM (2011)
20. Lokoč, J., Novák, D., Skopal, T., Sibirkina, N.: Thematic Web Images Collection. SIRET Research Group (2012), `http://siret.ms.mff.cuni.cz/twic`
21. Manjunath, B.S., Salembier, P., Sikora, T. (eds.): Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley & Sons, Inc., New York (2002)
22. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
23. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
24. MPEG-7. Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002 (2002)
25. Novak, D., Batko, M., Zezula, P.: Metric Index: An efficient and scalable solution for precise and approximate similarity search. Information Systems 36(4), 721–733 (2011)
26. Puzicha, J., Buhmann, J.M., Rubner, Y., Tomasi, C.: Empirical evaluation of dissimilarity measures for color and texture. In: Proc. IEEE International Conference on Computer Vision, vol. 2, pp. 1165–1172 (1999)
27. Skopal, T., Bartoš, T., Lokoč, J.: On (not) indexing quadratic form distance by metric access methods. In: Proc. Extending Database Technology (EDBT). ACM (2011)
28. Spyrou, E., Le Borgne, H., Mailis, T., Cooke, E., Avrithis, Y., O'Connor, N.: Fusing MPEG-7 visual descriptors for image classification, pp. 847–852 (September 2005)
29. Stanchev, P., Amato, G., Falchi, F., Gennaro, C., Rabitti, F., Savino, P.: Selection of MPEG-7 image features for improving image similarity search on specific data sets. In: 7th IASTED International Conference on Computer Graphics and Imaging, CGIM, pp. 395–400. Citeseer (2004)
30. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 947–963 (2001)