# CoPhIR Image Collection under the Microscope


## Abstract

*The Content-based Photo Image Retrieval (CoPhIR) dataset is the largest available database of digital images with corresponding visual descriptors. It contains five MPEG-7 global descriptors extracted from more than 106 million images from Flickr photo-sharing system. In this paper, we analyze this dataset focusing on 1) efficiency of similarity-based indexing and searching and on 2) expressiveness of combination of these descriptors with respect to subjective perception of visual similarity. We treat the descriptors as metric spaces and then combine them into a multi-metric space. We analyze distance distributions of individual descriptors, measure intrinsic dimensionality of these datasets and statistically evaluate correlation between these descriptors. Further, we use two methods to assess subjective accuracy and satisfaction of similarity retrieval based on a combination of descriptors that is recommended for CoPhIR, and we compare these results on databases of 10 and 100 million CoPhIR images. Finally, we suggest, explore and evaluate two approaches to improve the accuracy: 1) applying logarithms in order to weaken influence of a single descriptor contribution if it deviates from the rest, and 2) the possibility of categorization of the dataset and identifying visual characteristics important for individual categories.*


## 1 Introduction

With the rapid growth of the volume and diversity of digital data, the need of efficient storage and retrieval methods is indisputable. Since the classical databases are not suitable for many new data types (images, video, etc.), a new approach of similarity searching has been intensively researched in the recent years.

Nowadays, first Web-scale applications of similarity search begin to emerge. One of the largest is the MUFIN[1] image search system [9], that organizes data from CoPhIR[2] – a unique database providing visual descriptors for more than 100 million images collected from the Web. Use this public demonstration and other content-based systems over CoPhIR, we were able to gather statistics on user satisfaction and retrieval accuracy. Exploiting this data and studying the CoPhIR dataset itself, we explore ways to improve user satisfaction with content-based image search. In this paper, we present two approaches: we examine the combination of visual descriptors in multi-metric space, and we consider possible categorization of the dataset and seek individual similarity measures for each category.

### Objectives

Let us summarize the purpose and contribution of the paper:

- we introduce the CoPhIR dataset, in particular its five MPEG-7 visual descriptors (Section 2.1);
- we analyze distributions of individual descriptors spaces, their intrinsic dimensionalities and mutual relations between these descriptors (Sections 2.2, 2.3);
- we examine several possible aggregation functions and dispute their effect on both the user satisfaction and the search efficiency (Sections 3.1, 3.2);
- we explore the potency of image categorization and tuning search for individual categories (Section 3.3);
- all approaches discussed are evaluated via a user-satisfaction methodology recommended for MPEG-7.

### Related Work

The creation of the CoPhIR dataset has been studied by its authors within the SAPIR project [5]. The authors of [1] provide their experience with the five CoPhIR visual descriptors and they suggest a general combination function that is considered to be suitable for CoPhIR. The indexability of a general metric space was researched by several authors. Chavez and Navarro studied the concept of intrinsic dimensionality [2]. Skala [11] examined effect of various distance measures on intrinsic dimensionality and later [12] studies this and other properties for datasets from SISAP metric space library.

[1] Multi-Feature Indexing Network, http://mufin.fi.muni.cz/imgsearch/

[2] Content-based Photo Image Retrieval, http://cophir.isti.cnr.it

## 2 The CoPhIR Database

The CoPhIR dataset[3] [5] consists of metadata extracted from the Flickr photo sharing system[4]. The collection is composed mostly of outdoor and indoor photos, and there are also a few images of e-shops products, cartoon images, hand drawings, paintings, etc. The following information is stored for each image:

- link to corresponding entry at Flickr Web site;
- thumbnail image;
- user-specified metadata from the corresponding Flickr entry (title, GPS location, tags, comments, etc.);
- five MPEG-7 visual descriptors extracted from the image content (stored in XML format).

We focus on the content-based visual information and its utilization for indexing and searching the image database. MPEG-7 [7, 6] is a standard for description of multimedia content. It provides descriptors for various data types – audio, graphics, text, video, and scenario. The graphics descriptors are divided into groups focusing on color, texture and shape. The CoPhIR authors have selected five descriptors that seem to perform quite well on non-specified image databases [1]: Scalable Color, Color Structure, Color Layout, Edge Histogram and Homogeneous Texture.

### 2.1 CoPhIR Visual Descriptors

Let us describe the five MPEG-7 visual descriptors provided for each image in CoPhIR database. There is a function defined for each of the descriptors [6] that measures the *distance* (dissimilarity) between two instances of these descriptors (extracted from two images).

**Scalable Color** Scalable color descriptor is derived from a color histogram in the Hue-Saturation-Value color space with fixed space quantization. The histogram values are extracted, normalized and non-linearly mapped into a four-bit integer representation. Then the Haar transformation is applied which performs primitive low-pass and high-pass filters. In CoPhIR, the 64 coefficients version of this descriptor is used. The distance between two scalable color values is measured by the $L_1$ metric (sum of absolute differences).

**Color Structure** Color structure descriptor is also based on color histograms but aims at identifying localized color distributions using a small structuring matrix. Instead of considering each pixel separately, the extraction method embeds color information into the descriptor by taking into account all colors in a structuring element of $8 \times 8$ pixels that slides over the image. Unlike color histogram, 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. Again, the $L_1$ metric is used to compute descriptors distances.

**Color Layout** This descriptor is obtained by applying the Discrete cosine transform on a 2-D array (usually $8 \times 8$ blocks) of local representative colors in three color spaces (Y, Cb, and Cr). It was designed to efficiently represent spatial distribution of colors with no dependency on image format, resolution, and bit-depth. The 12 coefficients version of this descriptor is used in CoPhIR. The distance between two objects is computed as a sum of $L_2$ distances in each of the three color spaces.

**Edge Histogram** Edge histogram descriptor represents the local-edge distribution in the image. The image is subdivided into $4 \times 4$ sub-images and edges in each sub-image are categorized into five types: vertical, horizontal, $45^\circ$ diagonal, $135^\circ$ diagonal, and non-directional edges. Thus we get a vector of 80 coefficients (5 values for each of the 16 sub-images). Based on the descriptor values which represent local edge histograms, the semi-global and the global histograms can be computed. The distance between two edge histogram values is then computed as a sum of weighted sub-sums of absolute differences for each of the three histograms.

**Homogeneous Texture** This descriptor characterizes the region texture using the mean energy and the energy deviation from a set of 30 frequency channels. The extraction is done as follows: The image is first filtered with a bank of orientation and scale tuned filters (modeled using Gabor functions) using Gabor filters. The first and the second moments of the energy in the frequency domain in the corresponding sub-bands are then used as the components of the texture descriptor. The complete form of this descriptor consisting of 62 coefficients is used in CoPhIR (the overall mean and deviation of the image and the mean energy and deviation for each of the channels). This descriptor is not extracted for images smaller then $128 \times 128$ pixels.

### 2.2 Descriptors: Metric Space Properties

Let us study properties of the data spaces of the five MPEG-7 descriptors in CoPhIR dataset. We saw in the previous section that each of the MPEG-7 descriptors is accompanied with a distance function that is used to evaluate the distance (dissimilarity) of two feature vectors extracted from these images. All the distance measures are metric

[3] http://cophir.isti.cnr.it
[4] http://www.flickr.com

functions (see below) and therefore we model this data as metric spaces and study their metric-based space properties.

### Metric function

The metric space [13] is considered to be the most general data model for similarity search which can still be indexed and searched efficiently. The model treats the data as unstructured objects together with a function which measures proximity of object pairs. Formally, *metric space* $\mathcal{M}$ is a pair $\mathcal{M} = (\mathcal{D}, d)$, where $\mathcal{D}$ is the *domain* of objects and $d$ is the total *distance function* $d : \mathcal{D} \times \mathcal{D} \longrightarrow \mathbb{R}$ satisfying the following postulates for all objects $x, y, z \in \mathcal{D}$:

| | | |
|---|---|---|
| $d(x, x) = 0$ | *reflexivity* | (1) |
| $d(x, y) > 0$ | *strict positiveness* | (2) |
| $d(x, y) = d(y, x)$ | *symmetry* | (3) |
| $d(x, y) \leq d(x, z) + d(z, y)$ | *triangle inequality* | (4) |

The semantics of this concept is: The smaller the distance between two objects, the more similar they are. The metric space is typically searched by queries which follow the query-by-example paradigm. A query is formed by an *object* $q \in \mathcal{D}$ and some *constraint* on the data to be retrieved from the indexed dataset $X \subseteq \mathcal{D}$. There are two basic types of these queries: (1) the *range query* $R(q, r)$, which retrieves all objects $o \in X$ within the range $r$ from $q$ (i.e. $\{o|d(q, o) \leq r\}$), and (2) the *nearest neighbors query* $kNN(q, k)$, which returns the $k$ objects from $X$ with the smallest distances to $q$.

### Distance Histogram

The MPEG-7 specification defines minimal and maximal possible values for each descriptor. Thus we can compute the maximal possible distance between two objects under each of the respective descriptor distance measures. However, the actual distances of objects in dataset depend on the type of images in the collection (e.g. images from electron microscope have different properties than outdoor photos). To learn about actual distribution of object distances in the CoPhIR collection, we computed the five distances (one for each descriptor) of 500,000 random pairs of objects. Figure 1 shows this distance distribution as a *histogram of distances* [2] – $x$-axis shows the distance normalized by the maximal theoretical value and $y$-axis shows the numbers of distances that fall into a certain distance interval. As the actual $y$-axis values depend strictly on number of tested distances and quantization of the $x$-axis, the $y$-axis is not labeled by specific values.

### Intrinsic Dimensionality

Depending on the distance distribution, the metric space can be more or less difficult to search. Many authors denote this phenomenon as intrinsic dimensionality of the metric space [2]. The intrinsic dimensionality of a data space can be defined [2] as $\rho = \mu^2/(2 \cdot \sigma^2)$, where $\mu$ and $\sigma^2$ are the mean and variance of its histogram of distances. Metric spaces with lower values of intrinsic dimensionality are more suitable for indexing and searching.

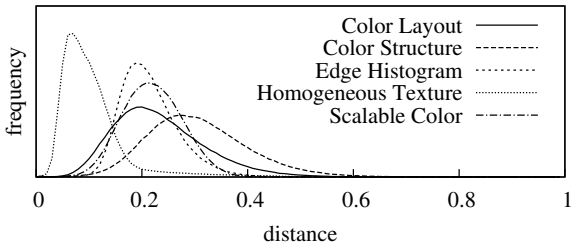


**Figure 1. Distance histograms of the five MPEG-7 descriptors.**

The following table shows the intrinsic dimensionalities of the five descriptor metric spaces.

| MPEG-7 descriptor | Intrinsic dimensionality |
|---|---|
| Color Structure | 5.116 |
| Color Layout | 3.576 |
| Edge Histogram | 7.507 |
| Homogeneous Texture | 1.323 |
| Scalable Color | 7.144 |

We can observe that the intrinsic dimensionality values correspond well with the distance histograms – the more narrow and high the histogram is, the worse the dimensionality and therefore searchability are. Intuitively, many objects with nearly the same distance imply that many distance computations must be evaluated to find the nearest ones (the objects cannot be simply eliminated using the triangle inequality property of the metric space).

## 2.3 Descriptors: Correlation

Having several image descriptors, it is natural to combine them in order to get more information about similarity between images. For this purpose, it could be favourable to uncover relationships between descriptors, for instance if two descriptors behave in the same way, there is no use in employing both of them. Therefore we used the 500,000 distances sample set to evaluate the statistical correlation of descriptors.

We can see from the results in Table 1 that there are some dependencies between the descriptors. As we expected, the color descriptors are correlated with each other more than with the texture descriptors and vice versa. The most correlated descriptor is Scalable Color, therefore it would be

| Descriptor | Color Layout | Color Structure | Edge Histogram | Hom. Texture | Scalable Color |
|---|---|---|---|---|---|
| **Color Layout** | 1.00 | **0.23** | 0.10 | 0.06 | **0.45** |
| **Color Structure** | **0.23** | 1.00 | **0.24** | 0.09 | **0.67** |
| **Edge Histogram** | 0.10 | **0.24** | 1.00 | **0.23** | 0.18 |
| **Hom. Texture** | 0.06 | 0.09 | **0.23** | 1.00 | 0.09 |
| **Scalable Color** | **0.45** | **0.67** | 0.18 | 0.09 | 1.00 |

**Table 1. Correlation between the MPEG-7 descriptors.**

the first candidate for omitting if we wanted to reduce the number of descriptors (to save space or processing time).

#### Normalization Problem

An interesting problem concerning descriptor combinations arises when we look closer at the distance histogram in Figure 1. All of the distances have been normalized, i.e. divided by the maximal possible value in order to obtain a number from $[0, 1]$. This is useful for observing and comparison of the results quality and it is essential for descriptors combination – influence of each descriptor on the aggregated distance should be the same (if we do not consider any weights). With a good normalization, it is also transparent for users to add weights to the descriptors according to their preferences.

However, we can see that the distributions of distances (mean values, variations) differ significantly in our histograms. Due to this fact, the influence of the descriptors is not equal. To solve this problem, we propose different normalization factors enumerated in the following table.

| Descriptor | Basic norm. | Equal norm. |
|---|---|---|
| Color Layout | 0.215 | 0.200 |
| Color Structure | 0.282 | 0.201 |
| Edge Histogram | 0.203 | 0.204 |
| Homogeneous Texture | 0.097 | 0.194 |
| Scalable Color | 0.204 | 0.201 |

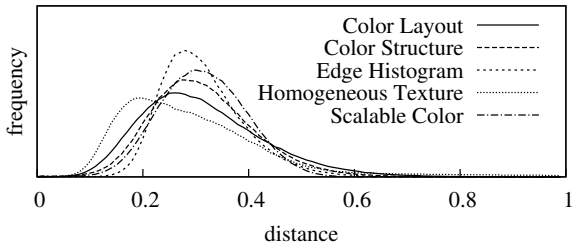


**Figure 2. Distance histograms of MPEG-7 descriptors with equal normalization.**

Exploiting our knowledge of the data distributions we these factors transform the distances in such way that they remain in interval of $[0, 1]$ but their distance distributions are more similar – see Figure 2. The overall characteristics of the metric space remain the same – the intrinsic dimensionality is not influenced by linear transformations.

## 3 Visual Descriptors Aggregation

Each of the MPEG-7 visual descriptors covers a different characteristic of an image, so their similarity measures can only compare images from a specific point of view. However, users are usually interested in overall similarity, i.e. a measure that combines the respective basic descriptors. In practice, we use an aggregation function that accepts the respective descriptors' distances as arguments and provides the overall similarity as a result.

#### Aggregation Function Properties

Let $m$ be the number of descriptors to combine. The *aggregation function* is defined as a function $t : [0, 1]^m \rightarrow [0, \infty)$, such that $t(x_1, \ldots, x_m) \leq t(x'_1, \ldots, x'_m)$ when $x_i \leq x'_i$ for each $i = 1, \ldots, m$. This property is called *monotonicity* and it is important, e.g. for the dynamic combining methods like the Threshold algorithm [4].

The overall similarity function is then defined as

$$D(x, y) = t(d_1(x, y), \ldots, d_m(x, y)),$$

where $d_i(x, y)$ is a distance function of the $i^{th}$ descriptor. Note that such a function need not necessarily be a metric. If the combination is to be used in a metric-based index, we must verify that the resulting similarity function satisfies the metric postulates.

#### User Satisfaction Survey

Different descriptor aggregations express different notions of similarity between the images, e.g. we can focus on image color similarity, on the layout or texture of the images. One of the objectives of this paper is to study effectiveness of search in the CoPhIR dataset from a user perspective.

We have chosen two methods of experimental evaluation of the results, the first of which is *user satisfaction survey*. We have prepared a Web page where a user is given a random image from the CoPhIR dataset with 30 most similar images returned by a given aggregation function. The user is then asked to specify their satisfaction with the results on a scale from 1 (best) to 6 (worst). We had over 30 users of both genders, various ages, professions, and computer skills, and we obtained over 800 feedback answers.

#### Subjective Evaluation of Retrieval Accuracy

The second method we adopt, is a retrieval accuracy measure [8] recommended for MPEG-7 visual descriptors and used in Chapter 12 of MPEG-7 book [6]. This measure is called *Average Normalized Modified Retrieval Rate* (ANMRR) and is based on the concept of *ground truth* (GT) – a set of images from the dataset that are visually similar to a given query image $q$. ANMRR measure ranks images from GT according to their rank in the $k$ most similar images to $q$, whereas images that appear on a position over a certain $k$ are considered to have rank $1.25 \cdot k$. Ranks of images from GT are averaged, normalized and, finally, retrieval accuracy ANMRR of a given approach is established as an average of these values for a set of query images $q$. ANMRR values range from 0 to 1 where 0 and 1 are assigned to the best and worst accurate methods, respectively.

In our case, we can hardly precisely find ground-truth images in sets of 10 or 100 million images. Thus, we adjust ANMRR so that we 1) fixate the value of $k$ for all queries ($k = 30$, in all the experiments), 2) we expect each GT set to have more than $k$ images, and 3) we consider only the "first $k$ images from the GT". In practice, this means that we manually point out ground-truth images (relevant images) within each answer (let us denote the number of these images $g : 0 \leq n_{gt} \leq k$) and we consider the rest $k - n_{gt}$ images from GT to be missed. This approach can slightly handicap queries with fewer actual ground-truth images in the dataset but it works precisely when we average the values over all queries and compare individual retrieval methods according to this average.

### 3.1 Weighted Sum Aggregation Function

A commonly used method for combining similarity measures is to add a weight multiplication to each of the measures and then sum them. More formally

$$t(x_1, ..., x_m) = w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_m \cdot x_m,$$

where $w_i \in \mathbb{R}^+$. Additionally, we can divide the result by the sum of the weights to acquire a normalized measure interval of $[0, 1]$.

We can easily prove that the weighted sum of metric functions is also a metric function. First, the multiplication by a positive weight constant does not break the metric properties, i.e. $\forall c \in \mathbb{R}^+ : D(x, y) = c \cdot d(x, y)$ is a metric. The result is zero if and only if the original distance was zero (1), otherwise the result is positive (2). Symmetry is inherited from the original metric (3) and the triangle inequality holds since multiplication by a positive number keeps the equation intact (4).

Second, the sum of metric functions (on the same objects) is also a metric function, i.e. $D(x, y) = \sum_i d_i(x, y)$ is a metric. Since all $d_i$ functions are metric, the result is always either greater or equal to zero (2). The result can be zero if and only if all the $d_i$ distances are zero (1). Since the $d_i$ functions are symmetric, swapping the arguments gives the same result (3). Triangular inequality can be proved by applying the induction using $d_i(x, y) + d_{i+1}(x, y) + d_i(y, z) + d_{i+1}(y, z) \leq d_i(x, z) + d_{i+1}(x, z)$, which holds due to the fact that $0 \leq d_i(x, y) + d_i(y, z) - d_i(x, z)$ which comes from the triangular inequality of function $d_i$.

#### CoPhIR Defined Weighted Sum

Amato et al. suggested a combination function [1] that was slightly modified and used in the MUFIN system. The descriptors are aggregated by a weighted sum – the respective weights are summarized in the following table.

| MPEG-7 descriptor | Weight |
|---|---|
| Scalable Color | 2.5 |
| Color Structure | 2.5 |
| Color Layout | 1.5 |
| Edge Histogram | 4.5 |
| Homogeneous Texture | 0.5 |

The distance histogram of the CoPhIR dataset for this overall function is shown in Figure 3. As expected, the distribution of distances is normal, with the mean value slightly above 0.2. This results in the intrinsic dimensionality 12.9, which means that this function is rather difficult to index.

We have tested subjective accuracy of this similarity function as specified at the beginning of this section. We

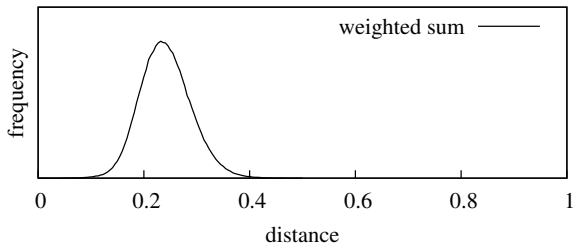


**Figure 3. Distance histogram for the overall CoPhIR aggregation function.**


**Figure 4. Results of standard combination function of 10 and 100 million CoPhIR datasets.**

used two subsets of the CoPhIR dataset – 10 million images and 100 million images. In both cases, users were given query images from a randomly selected set of 50 images. The results of the "user satisfaction survey" are shown in the following table. The satisfaction column is an average of the user-selected values from 1 (best satisfaction) to 6 (worst satisfaction) and we also show the total number of queries answered and the number of unique users.

| Dataset | Satisfaction | Queries | Users |
|---|---|---|---|
| 10M | 2.63 | 446 | 31 |
| 100M | 2.17 | 401 | 28 |

We can conclude that users were clearly satisfied with the results and, as expected, the bigger dataset yields a higher satisfaction. However, there is still a space for improvement, e.g. we had about 17 answers marked as 6 (completely unsatisfactory).

The second evaluation was conducted using the ANMRR methodology on the same set of query images. Let us recall that this is based on identification of the ground truth for each query image and the ANMRR values range from 0 (full $k$ of ground-truth images returned for all queries) to 1 (no ground-truth image retrieved for any query). An example of such results is in Figure 4 – the first result (on a 10 million database) contains five ground-truth images (pictures of T-shirts) and the second contains eight.

The results, summarized in the following table, are in correspondence with the first experiment. We can observe that the accuracy improved noticeably when the dataset size increased. Let us realize that we have adopted presumption that "the ground truth of a query image is at least $k$ ($k = 30$)", which does not necessarily hold for all queries, and thus it is impossible to reach ANMRR near zero.

| Dataset | ANMRR | # improved | # worsen |
|---|---|---|---|
| 10M | 0.49 | - | - |
| 100M | 0.41 | 62 % | 28 % |

The table also reports on the number of queries for which the result has improved/worsened on the 100 million dataset (the remaining 10 % query results had approximately equal accuracy). We can see that majority query images exhibit improvement in retrieval accuracy but, quite surprisingly, the subjective quality was slightly worsened for about 28 % images. For these images, enlargement of the database introduced images that are closer according to the CoPhIR similarity function but they "pushed out" some subjectively better result images. Also, approximation adopted by MUFIN [9] could exhibit a bit worse recall in specific cases.

### 3.2 Logarithmic Aggregation Function

During analysis of the user satisfaction data, we have observed that some of the user-preferred images received higher distances because of just one descriptor. When we removed such descriptor from the aggregation, we received better results (from user satisfaction point of view) for that particular image. This inspired us to try to neglect the differences of higher distances by using logarithms. More formally, we used an aggregation function

$$t(x_1, ..., x_m) = w_1 \cdot \log_{b_1}(x_1+1) + \cdots + w_m \cdot \log_{b_m}(x_m+1),$$

where $b_i \in (1, \infty)$ and $w_i \in \mathbb{R}^+$. The function can be normalized by dividing the result by a constant $c = w_1 \cdot \log_{b_1}(1+1) + \cdots + w_m \cdot \log_{b_m}(1+1)$.

Since we have shown that a weighted sum of metric functions is a metric function, we only need to show that $D(x, y) = \log_b(d(x, y) + 1)$ is a metric function too. Function $D$ is always positive (observe the addition of 1) and it is equal to zero if and only if the $d(x, y)$ is zero (1), (2). Symmetry is inherited from $d$ (3). The triangular inequality can be proved by applying logarithm (which is monotonically increasing for bases greater than 1) to $d(x, y) + d(y, z) + d(x, y) \cdot d(y, z) + 1 \geq d(x, z) + 1$. The equation comes directly from the triangular inequality and positiveness properties of $d$. The proof (4) is then finished by a few straightforward derivations of the equation's left side.

**Experimental Results**

We applied natural logarithm on all five descriptors and we used the same weights as specified in Section 3.1. Figure 5 shows the new histogram of distances. We can observe that, comparing to standard CoPhIR overall function, the distances are more condensed around slightly higher mean value, which is worse for the indexing. This is confirmed by higher intrinsic dimensionality, value of which is now 16.2.

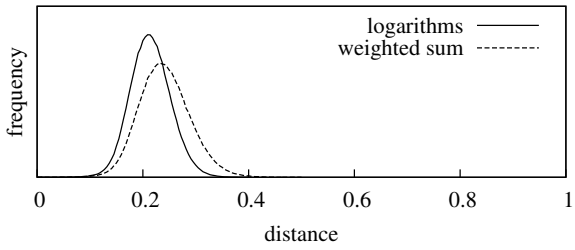


**Figure 5. Distance histogram for logarithms.**

Repeating the first user-satisfaction experiment for the logarithms function was impossible because of a limited availability of the testing group and thus we report only on the ANMRR values in the table below.

| Dataset | ANMRR | # improved | # worsen |
|---|---|---|---|
| 10M | 0.49 | - | - |
| logarithms | 0.43 | 55 % | 25 % |

We can see that the accuracy of the results has improved, in general. The improvement is countered by worsened intrinsic dimensionality and also by increased computational demands of the aggregation function with logarithms.

## 3.3 Image Categorization

The subject of automatic image categorization and labeling according to human visual perception has been studied extensively both in the field of psychology and computer science (see Section 3.3 of [3] for a comprehensive survey). "Designing representations of human visual categories is difficult because category instances are structurally variable. (...) The variability persists at different levels ranging from the single contour to the entire configuration of features or parts." [10]

Our aim is not to introduce a novel approach to automatic categorization but to study how categorization of CoPhIR database could improve the similarity retrieval according to the visual descriptors provided by CoPhIR. Namely, we report on the following experiment:

- we categorize the set of query images used above;
- for each image category, we try to identify characteristics important for images from this category and tune descriptors' weights for the category;
- we evaluate and compare retrieval accuracy using the global weights and the new weights combinations.

We believe that although identifying common visual concepts in an image category is very complicated, identifying *important features* is a different and possibly feasible task.

**Query Categorization** One of standard image databases widely used in CBIR is the Corel photo collection. It provides images categorized into 600 classes. Recently, Rasche [10] has comprehensively studied categorization of this dataset and built a category tree. He filters out 240 classes that are not based on visual perception of the images and he groups the rest 360 categories into 112 *basic-level categories*. These are further grouped into ten top-level categories. We adopt this concept and manually categorize the set of query objects into the ten top-level categories – see their list and their percentage numbers in Table 2 (the first two lines). We have added category *drawings* because the original ten are purely photograph categories. The partitioning is not uniform and follows distribution of CoPhIR (the queries were selected randomly uniformly).

**Weights for Categories** Let us continue in the user-satisfaction experiment described in previous sections. The third row of Table 2 shows the ANMRR values for the 10M dataset averaged over individual query image categories. Considering only categories having at least 5 % of the query set (the bold values), we can see that ANMRRs range from 0.23 to 0.61. We focused on the six categories with ANMRR over 0.3 and tried to individually tune descriptors weights to suit to query objects in these categories. The experiment was performed on a distributed memory-based 10M storage with a sequential-scan and adjustable distance function (one query evaluation took about 17 s on 16 CPUs). We experimentally identified the following rules that seem to improve the search quality for the query categories.

| category | weights modification |
|---|---|
| buildings | Edge Histogram (EH) = 9 |
| landscapes | Color Layout (CL) = 3.5 |
| parts | EH = 6, CL = 3.5 |
| persons | EH = 6, CL = 0, H. Texture (HT) = 0 |
| vehicles | EH = 9, HT = 0 |
| drawings | EH = 6, CL = 3, HT = 0 |

**Retrieval Accuracy Evaluation** We have tuned the weights for individual categories regarding always half of the queries from given category. The ANMRR accuracy evaluation was then done on the entire query set. The last row in Table 2 shows ANMRR results for the new weights as stated above. We can see improvement for all considered

| **category** | activities | animals | buildings | food | landscapes | parts | persons | plants | textures | vehicles | drawings |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **number** | 2% | 2% | 10% | 4% | 20% | 12% | 24% | 10% | 2% | 8% | 6% |
| **ANMRR** | 0.67 | 0.94 | **0.56** | 0.67 | **0.44** | **0.61** | **0.49** | **0.23** | 0.68 | **0.42** | **0.57** |
| **tuned** | | | 0.45 | | 0.38 | 0.55 | 0.42 | | | 0.38 | 0.49 |

**Table 2. Ground truth-based user satisfaction (ANMRR) for categorized query images.**

categories (the improvement varies between 0.04 and 0.11). The overall influence can be summarized as follows:

| **Dataset** | **ANMRR** | **# improved** | **# worsen** |
|---|---|---|---|
| 10M | 0.49 | - | - |
| categorized | 0.43 | 67.5 % | 10 % |

The categorization worsened the retrieval quality only for 10 % of the influenced query images. The ANMRR improvement down to 0.43 is comparable with the shift reached by multiplying the dataset to 100M (ANMRR = 0.41). Again, let us realize that the ANMRR values are to be compared mutually and it is probably impossible to reach ANMRR near zero (see Section 3.1).

## 4 Conclusion

Formation of the CoPhIR dataset enabled occurrence of real Web-scale systems for image content-based retrieval. A chase after increasing the data volume is certainly meaningful because the subjective feeling from a general image search begins to be satisfying only when the database reaches a level of at least tens of millions images. In this work, we tried to capture subjective impression by both an unspecified "satisfaction" experiment with dozens of users and by a more rigorous approach based on a query *ground truth*. We performed this testing on 10 and 100 million image databases and compared the results. We observed expected improvement in retrieval accuracy for majority of the query images but, quite surprisingly, the subjective quality was slightly worsened for about 30 % of the tested images. For these specific images, enlargement of the database introduced objects that are closer according to the similarity function but are subjectively less similar.

We suggest and prove a modification of the descriptors-aggregation function by applying logarithms on individual visual descriptors. This seems to improve the result quality by eliminating cases when an image is excluded from the query result only because a single descriptors. We also categorize the query images and seek a specific weighted descriptor combination for individual categories. This approach seems to be very promising and automatic categorization of the dataset could certainly improve user satisfaction with image content-based retrieval. Finally, we studied the data space properties of the MPEG-7 descriptors – distance distribution, intrinsic dimensionality, and also statistical correlation of individual descriptors.

## 5 Acknowledgments

This work has been supported by EU IST FP6 project 045128 (SAPIR), the national research project 1ET100300419, and the Czech Grant Agency project 201/07/P240.

## References

[1] G. Amato, F. Falchi, C. Gennaro, F. Rabitti, P. Savino, and P. Stanchev. Improving image similarity search effectiveness in a multimedia content management system. In *Proc. of Workshop on Multimedia Information System (MIS)*, pages 139–146, 2004.

[2] E. Chvez and G. Navarro. Measuring the dimensionality of general metric spaces. Technical Report TR/DCC-00-1, Department of Computer Science, University of Chile, 2000.

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

[4] R. Fagin. Combining fuzzy information: an overview. *SIGMOD Record*, 31:2002, 2002.

[5] F. Falchi, C. Lucchese, R. Perego, and F. Rabitti. CoPhIR: Content-based photo image retrieval, May 2008. http://cophir.isti.cnr.it/CoPhIR.pdf.

[6] B. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[7] MPEG-7. Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002, 2002.

[8] P. Ndjiki-Nya et al. Subjective evaluation of the MPEG-7 retrieval accuracy measure (ANMRR). Technical Report ISO/IEC JTC1/SC29/WG11 (MPEG) doc. M6029, Geneva, Switzerland, May 2000.

[9] D. Novak, M. Batko, and P. Zezula. Web-scale system for image similarity search: When the dreams are coming true. In *Proceedings of CBMI 2008*. IEEE, 2008.

[10] C. Rasche. An approach to the parameterization of structure for fast categorization. 2009. To appear. http://www.allpsych.uni-giessen.de/rasche/research /res_img_classification.htm.

[11] M. Skala. Measuring the difficulty of distance-based indexing. In *Proceedings of SPIRE 2005, Argentina, 2005*, volume 3772 of *LNCS*, pages 103–114. Springer, 2005.

[12] M. Skala. Counting distance permutations. *Journal of Discrete Algorithms*, 7(1):49–61, 2009.

[13] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer-Verlag, 2006.