Distance Measures for MPEG-7-based Retrieval

Horst Eidenberger

Vienna University of Technology, Institute of Software Technology and Interactive Systems
Favoritenstrasse 9-11 – A-1040 Vienna, Austria
Tel. + 43-1-58801-18853

eidenberger@ims.tuwien.ac.at

ABSTRACT

In visual information retrieval the careful choice of suitable proximity measures is a crucial success factor. The evaluation presented in this paper aims at showing that the distance measures suggested by the MPEG-7 group for the visual descriptors can be beaten by general-purpose measures. Eight visual MPEG-7 descriptors were selected and 38 distance measures implemented. Three media collections were created and assessed, performance indicators developed and more than 22500 tests performed. Additionally, a quantisation model was developed to be able to use predicate-based distance measures on continuous data as well. The evaluation shows that the distance measures recommended in the MPEG-7-standard are among the best but that other measures perform even better.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Query formulation*, *Retrieval models.*

General Terms

Algorithms, Measurement, Experimentation, Performance, Theory.

Keywords

Visual Information Retrieval, Content-based Image Retrieval, Content-based Video Retrieval, Similarity Measurement, Distance Measurement, Similarity Perception, MPEG-7.

1. INTRODUCTION

The MPEG-7 standard defines – among others – a set of descriptors for visual media. Each descriptor consists of a feature extraction mechanism, a description (in binary and XML format) and guidelines that define how to apply the descriptor on different kinds of media (e.g. on temporal media). The MPEG-7 descriptors have been carefully designed to meet – partially complementary – requirements of different application domains: archival, browsing, retrieval, etc. [9]. In the following, we will exclusively deal with the *visual* MPEG-7 descriptors in the context of media *retrieval*.

MIR'03, November 7, 2003, Berkeley, California, USA.

The visual MPEG-7 descriptors fall in five groups: colour, texture, shape, motion and others (e.g. face description) and sum up to 16 basic descriptors. For retrieval applications, a rule for each descriptor is mandatory that defines how to measure the similarity of two descriptions. Common rules are distance functions, like the Euclidean distance and the Mahalanobis distance. Unfortunately, the MPEG-7 standard does not include distance measures in the normative part, because it was not designed to be (and should not exclusively understood to be) retrieval-specific. However, the MPEG-7 authors give recommendations, which distance measure to use on a particular descriptor. These recommendations are based on accurate knowledge of the descriptors' behaviour and the description structures.

In the present study a large number of successful distance measures from different areas (statistics, psychology, medicine, social and economic sciences, etc.) were implemented and applied on MPEG-7 data vectors to *verify* whether or not the recommended MPEG-7 distance measures are really the best for any reasonable class of media objects. From the MPEG-7 tests and the recommendations it does not become clear, how many and which distance measures have been tested on the visual descriptors and the MPEG-7 test datasets. The hypothesis is that analytically derived distance measures may be good in general but only a quantitative analysis is capable to identify the *best* distance measure for a specific feature extraction method.

The paper is organised as follows. Section 2 gives a minimum of background information on the MPEG-7 descriptors and distance measurement in visual information retrieval (VIR, see [3], [16]). Section 3 gives an overview over the implemented distance measures. Section 4 describes the test setup, including the test data and the implemented evaluation methods. Finally, Section 5 presents the results per descriptor and over all descriptors.

2. BACKGROUND

2.1 MPEG-7: visual descriptors

The visual part of the MPEG-7 standard defines several descriptors. Not all of them are really descriptors in the sense that they extract properties from visual media. Some of them are just structures for descriptor aggregation or localisation. The basic descriptors are Color Layout, Color Structure, Dominant Color, Scalable Color, Edge Histogram, Homogeneous Texture, Texture Browsing, Region-based Shape, Contour-based Shape, Camera Motion, Parametric Motion and Motion Activity.

Other descriptors are based on low-level descriptors or semantic information: Group-of-Frames/Group-of-Pictures Color (based on

Scalable Color), Shape 3D (based on 3D mesh information), Motion Trajectory (based on object segmentation) and Face Recognition (based on face extraction).

Descriptors for spatiotemporal aggregation and localisation are: Spatial 2D Coordinates, Grid Layout, Region Locator (spatial), Time Series, Temporal Interpolation (temporal) and SpatioTemporal Locator (combined). Finally, other structures exist for colour spaces, colour quantisation and multiple 2D views of 3D objects.

These additional structures allow combining the basic descriptors in multiple ways and on different levels. But they do not change the *characteristics* of the extracted information. Consequently, structures for aggregation and localisation were not considered in the work described in this paper.

2.2 Similarity measurement on visual data

Generally, similarity measurement on visual information aims at imitating human visual similarity perception. Unfortunately, human perception is much more complex than any of the existing similarity models (it includes perception, recognition and subjectivity).

The common approach in visual information retrieval is measuring *dis*-similarity as *distance*. Both, query object and candidate object are represented by their corresponding feature vectors. The distance between these objects is measured by computing the distance between the two vectors. Consequently, the process is independent of the employed querying paradigm (e.g. query by example). The query object may be natural (e.g. a real object) or artificial (e.g. properties of a group of objects).

Goal of the measurement process is to express a relationship between the two objects by their distance. Iteration for multiple candidates allows then to define a partial order over the candidates and to address those in a (to be defined) neighbourhood being *similar* to the query object. At this point, it has to be mentioned that in a multi-descriptor environment – especially in MPEG-7 – we are only half way towards a statement on similarity. If multiple descriptors are used (e.g. a descriptor scheme), a rule has to be defined how to combine all distances to a global value for each object. Still, distance measurement is the most important first step in similarity measurement.

Obviously, the main task of good distance measures is to *reorganise* descriptor space in a way that media objects with the highest similarity are nearest to the query object. If distance is defined minimal, the query object is always in the origin of distance space and similar candidates should form clusters around the origin that are as large as possible. Consequently, many well known distance measures are based on geometric assumptions of descriptor space (e.g. Euclidean distance is based on the metric axioms). Unfortunately, these measures do not fit ideally with human similarity perception (e.g. due to human subjectivity). To overcome this shortage, researchers from different areas have developed alternative models that are mostly predicate-based (descriptors are assumed to contain just binary elements, e.g. Tversky's Feature Contrast Model [17]) and fit better with human perception. In the following distance measures of both groups of approaches will be considered.

3. DISTANCE MEASURES

The distance measures used in this work have been collected from various areas (Subsection 3.1). Because they work on differently quantised data, Subsection 3.2 sketches a model for unification on the basis of quantitative descriptions. Finally, Subsection 3.3 introduces the distance measures as well as their origin and the idea they implement.

3.1 Sources

Distance measurement is used in many research areas such as psychology, sociology (e.g. comparing test results), medicine (e.g. comparing parameters of test persons), economics (e.g. comparing balance sheet ratios), etc. Naturally, the character of data available in these areas differs significantly. Essentially, there are two extreme cases of data vectors (and distance measures): predicate-based (all vector elements are binary, e.g. {0, 1}) and quantitative (all vector elements are continuous, e.g. [0, 1]).

Predicates express the *existence* of properties and represent high-level information while quantitative values can be used to measure and mostly represent low-level information. Predicates are often employed in psychology, sociology and other human-related sciences and most predicate-based distance measures were therefore developed in these areas. Descriptions in visual information retrieval are nearly ever (if they do not integrate semantic information) quantitative. Consequently, mostly quantitative distance measures are used in visual information retrieval.

The goal of this work is to compare the MPEG-7 distance measures with the most powerful distance measures developed in other areas. Since MPEG-7 descriptions are purely quantitative but some of the most sophisticated distance measures are defined exclusively on predicates, a model is mandatory that allows the application of predicate-based distance measures on quantitative data. The model developed for this purpose is presented in the next section.

3.2 Quantisation model

The goal of the quantisation model is to redefine the set operators that are usually used in predicate-based distance measures on continuous data. The first in visual information retrieval to follow this approach were Santini and Jain, who tried to apply Tversky's Feature Contrast Model [17] to content-based image retrieval [12], [13]. They interpreted continuous data as fuzzy predicates and used fuzzy set operators. Unfortunately, their model suffered from several shortcomings they described in [12], [13] (for example, the quantitative model worked only for one specific version of the original predicate-based measure).

The main idea of the presented quantisation model is that set operators are replaced by *statistical* functions. In [5] the authors could show that this interpretation of set operators is reasonable.

The model offers a solution for the descriptors considered in the evaluation. It is not specific to one distance measure, but can be applied to any predicate-based measure. Below, it will be shown that the model does not only work for predicate data but for quantitative data as well. Each measure implementing the model can be used as a substitute for the original predicate-based measure.

Generally, binary properties of two objects (e.g. media objects) can exist in both objects (denoted as a), in just one (b, c) or in none of them (d). The operator needed for these relationships are *UNION*, *MINUS* and *NOT*. In the quantisation model they are replaced as follows (see [5] for further details).

$$a = X_i \cap X_j = \sum_k s_k,\ s_k = \begin{cases} \frac{x_{ik} + x_{jk}}{2} & if\ M - \frac{x_{ik} + x_{jk}}{2} \le \varepsilon_1 \\ 0 & else \end{cases}$$

$$b = X_i - X_j = \sum_k s_k,\ s_k = \begin{cases} x_{ik} - x_{jk} & if\ M - (x_{ik} - x_{jk}) \le \varepsilon_2 \\ 0 & else \end{cases}$$

$$c = X_j - X_i = \sum_k s_k,\ s_k = \begin{cases} x_{jk} - x_{ik} & if\ M - (x_{jk} - x_{ik}) \le \varepsilon_2 \\ 0 & else \end{cases}$$

$$d = \neg X_i \cap \neg X_j = \sum_k s_k,\ s_k = \begin{cases} M - \frac{x_{ik} + x_{jk}}{2} & if\ \frac{x_{ik} + x_{jk}}{2} \le \varepsilon_1 \\ 0 & else \end{cases}$$

with:

$$X_i = (x_{ik})\ with\ x_{ik} \in [x_{\min}, x_{\max}]$$

$$M = x_{\max} - x_{\min}$$

$$\varepsilon_1 = \begin{cases} M\left(1 - \frac{\mu}{p}\right) & if\ p \ge \mu \\ 0 & else \end{cases} \quad where\ \mu = \frac{\sum_i \sum_k x_{ik}}{i.k}$$

$$\varepsilon_2 = \begin{cases} M\left(1 - \frac{\sigma}{p}\right) & if\ p \ge \sigma \\ 0 & else \end{cases} \quad where\ \sigma = \sqrt{\frac{\sum_i \sum_k (\mu - x_{ik})^2}{i.k}}$$

$$p \in R^+ \setminus \{0\}$$

a selects properties that are present in both data vectors (X_i, X_j representing media objects), *b* and *c* select properties that are present in just one of them and *d* selects properties that are present in neither of the two data vectors. Every property is selected by the *extent* to which it is present (*a* and *d*: mean, *b* and *c*: difference) and only if the amount to which it is present exceeds a certain threshold (depending on the mean and standard deviation over all elements of descriptor space).

The implementation of these operators is based on one assumption. It is assumed that vector elements measure on interval scale. That means, each element expresses that the measured property is "more or less" present ("*0*": not at all, "*M*": fully present). This is true for most visual descriptors and all MPEG-7 descriptors. A natural origin as it is assumed here ("*0*") is not needed.

Introducing *p* (called discriminance-defining parameter) for the thresholds $\varepsilon_1, \varepsilon_2$ has the positive consequence that *a, b, c, d* can then be controlled through a *single* parameter. *p* is an additional criterion for the behaviour of a distance measure and determines the thresholds used in the operators. It expresses how accurate data items are present (quantisation) and consequently, how accurate they should be investigated. *p* can be set by the user or automatically. Interesting are the limits:

$$1.\ p \to \infty \Rightarrow \varepsilon_1, \varepsilon_2 \to M$$

In this case, all elements (=properties) are assumed to be continuous (high quantisation). In consequence, all properties of a descriptor are used by the operators. Then, the distance measure is *not* discriminant for properties.

$$2.\ p \to 0 \Rightarrow \varepsilon_1, \varepsilon_2 \to 0$$

In this case, all properties are assumed to be predicates. In consequence, only binary elements (=predicates) are used by the operators (1-bit quantisation). The distance measure is then highly discriminant for properties.

Table 1. Quantisation model on predicate vectors.

X_i	X_j	*a*	*b*	*c*	*d*
(1)	(1)	1	0	0	0
(1)	(0)	0	1	0	0
(0)	(1)	0	0	1	0
(0)	(0)	0	0	0	1

Between these limits, a distance measure that uses the quantisation model is – depending on *p* – more or less discriminant for properties. This means, it selects a subset of all available description vector elements for distance measurement.

For both predicate data and quantitative data it can be shown that the quantisation model is reasonable. If description vectors consist of binary elements only, *p* should be used as follows (for example, *p* can easily be set automatically):

$$p \to 0 \Rightarrow \varepsilon_1, \varepsilon_2 = 0, e.g.\ p = \min(\mu, \sigma)$$

In this case, *a*, *b*, *c*, *d* measure like the set operators they replace. For example, Table 1 shows their behaviour for two one-dimensional feature vectors X_i and X_j. As can be seen, the statistical measures work like set operators. Actually, the quantisation model works accurate on predicate data for any $p \neq \infty$.

To show that the model is reasonable for quantitative data the following fact is used. It is easy to show that for predicate data some quantitative distance measures degenerate to predicate-based measures. For example, the L^1 metric (Manhattan metric) degenerates to the Hamming distance (from [9], without weights):

$$L^1 = \sum_k |x_{ik} - x_{jk}| \equiv b + c = Hamming\ distance$$

If it can be shown that the quantisation model is able to *reconstruct* the quantitative measure from the degenerated predicate-based measure, the model is obviously able to *extend* predicate-based measures to the quantitative domain. This is easy to illustrate. For purely quantitative feature vectors, *p* should be used as follows (again, *p* can easily be set automatically):

$$p \to \infty \Rightarrow \varepsilon_1, \varepsilon_2 = 1$$

Then, *a* and *d* become continuous functions:

$$M - \frac{x_{ik} + x_{jk}}{2} \le M \equiv true \Rightarrow a = \sum_k s_k\ where\ s_k = \frac{x_{ik} + x_{jk}}{2}$$

$$\frac{x_{ik} + x_{jk}}{2} \le M \equiv true \Rightarrow d = \sum_k s_k\ where\ s_k = M - \frac{x_{ik} + x_{jk}}{2}$$

b and *c* can be made continuous for the following expressions:

$$M - (x_{ik} - x_{jk}) \le M \equiv x_{ik} - x_{jk} \ge 0$$

$$\Rightarrow b = \sum_k s_k\ where\ s_k = \begin{cases} x_{ik} - x_{jk} & if\ x_{ik} - x_{jk} \ge 0 \\ 0 & else \end{cases}$$

$$M - (x_{jk} - x_{ik}) \le M \equiv x_{jk} - x_{ik} \ge 0$$

$$\Rightarrow c = \sum_k s_k\ where\ s_k = \begin{cases} x_{jk} - x_{ik} & if\ x_{jk} - x_{ik} \ge 0 \\ 0 & else \end{cases}$$

$$\Rightarrow b + c = \sum_k s_k\ where\ s_k = |x_{ik} - x_{jk}|$$

Table 2. Predicate-based distance measures.

No.	Measure	Comment
P1	$a - \alpha.b - \beta.c$	Feature Contrast Model, Tversky 1977 [17]
P2	a	No. of co-occurrences
P3	$b + c$	Hamming distance
P4	$\frac{a}{K}$	Russel 1940 [14]
P5	$\frac{a}{b+c}$	Kulczvnski 1927 [14]
P6	$\frac{bc}{K^2}$	Pattern difference [14]
P7	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Pearson 1926 [11]

$$b - c = \sum_k s_k \ \text{where} \ s_k = x_{ik} - x_{jk}$$

$$c - b = \sum_k s_k \ \text{where} \ s_k = x_{jk} - x_{ik}$$

This means, for sufficiently high p every predicate-based distance measure that is either not using b and c or just as $b+c$, $b-c$ or $c-b$, can be transformed into a continuous quantitative distance measure. For example, the Hamming distance (again, without weights):

$$b + c = \sum_k s_k \ \text{where} \ s_k = \left| x_{ik} - x_{jk} \right| = \sum_k \left| x_{ik} - x_{jk} \right| = L^1$$

The quantisation model successfully reconstructs the L^1 metric and no distance measure-specific modification has to be made to the model. This demonstrates that the model is reasonable. In the following it will be used to extend successful predicate-based distance measures on the quantitative domain.

The major advantages of the quantisation model are: (1) it is application domain independent, (2) the implementation is straightforward, (3) the model is easy to use and finally, (4) the new parameter p allows to control the similarity measurement process in a new way (discriminance on property level).

3.3 Implemented measures

For the evaluation described in this work next to predicate-based (based on the quantisation model) and quantitative measures, the distance measures recommended in the MPEG-7 standard were implemented (all together 38 different distance measures).

Table 2 summarises those predicate-based measures that performed best in the evaluation (in sum 20 predicate-based measures were investigated). For these measures, K is the number of predicates in the data vectors X_i and X_j. In P1, the *sum* is used for Tversky's *f()* (as Tversky himself does in [17]) and α, β are weights for element b and c. In [5] the author's investigated Tversky's Feature Contrast Model and found $\alpha=1, \beta=0$ to be the optimum parameters.

Some of the predicate-based measures are very simple (e.g. P2, P4) but have been heavily exploited in psychological research. Pattern difference (P6) – a very powerful measure – is used in the statistics package SPSS for cluster analysis. P7 is a correlation coefficient for predicates developed by Pearson.

Table 3 shows the best quantitative distance measures that were used. Q1 and Q2 are metric-based and were implemented as representatives for the entire group of Minkowski distances. The w_i are weights. In Q5, μ_i, σ_i are mean and standard deviation for the elements of descriptor X_i. In Q6, m is $\frac{M}{2}$ (=0.5). Q3, the Canberra metric, is a normalised form of Q1. Similarly, Q4, Clark's divergence coefficient is a normalised version of Q2. Q6 is a further-developed correlation coefficient that is invariant against sign changes. This measure is used even though its particular properties are of minor importance for this application domain. Finally, Q8 is a measure that takes the differences between adjacent vector elements into account. This makes it structurally different from all other measures.

Obviously, one important distance measure is missing. The Mahalanobis distance was not considered, because different descriptors would require different covariance matrices and for some descriptors it is simply impossible to define a covariance matrix. If the identity matrix was used in this case, the Mahalanobis distance would degenerate to a Minkowski distance.

Additionally, the recommended MPEG-7 distances were implemented with the following parameters: In the distance measure of the Color Layout descriptor all weights were set to "1" (as in all other implemented measures). In the distance measure of the Dominant Color descriptor the following parameters were used: $w_1 = 0.7, w_2 = 0.3, \alpha = 1, T_d = 20$ (as recommended). In the Homogeneous Texture descriptor's distance all $\alpha(k)$ were set to "1" and matching was done rotation- and scale-invariant.

Important! Some of the measures presented in this section are *distance* measures while others are *similarity* measures. For the tests, it is important to notice, that all similarity measures were *inverted* to distance measures.

4. TEST SETUP

Subsection 4.1 describes the descriptors (including parameters) and the collections (including ground truth information) that were used in the evaluation. Subsection 4.2 discusses the evaluation method that was implemented and Subsection 4.3 sketches the test environment used for the evaluation process.

4.1 Test data

For the evaluation eight MPEG-7 descriptors were used. All colour descriptors: Color Layout, Color Structure, Dominant Color, Scalable Color, all texture descriptors: Edge Histogram, Homogeneous Texture, Texture Browsing and one shape descriptor: Region-based Shape. Texture Browsing was used even though the MPEG-7 standard suggests that it is not suitable for retrieval. The other basic shape descriptor, Contour-based Shape, was not used, because it produces structurally different descriptions that cannot be transformed to data vectors with elements measuring on interval-scales. The motion descriptors were not used, because they integrate the temporal dimension of visual media and would only be comparable, if the basic colour, texture and shape descriptors would be aggregated over time. This was not done. Finally, no high-level descriptors were used (Localisation, Face Recognition, etc., see Subsection 2.1), because – to the author's opinion – the behaviour of the basic descriptors on elementary media objects should be evaluated *before* conclusions on aggregated structures can be drawn.

Table 3. Quantitative distance measures.

No.	Measure	Comment	No.	Measure	Comment
Q1	$\sum_k w_i \left\| x_{ik} - x_{jk} \right\|$	City block distance (L^1)	Q2	$\sqrt{\sum_k w_i \left(x_{ik} - x_{jk} \right)^2}$	Euclidean distance (L^2)
Q3	$\sum_k \frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}}$	Canberra metric, Lance, Williams 1967 [8]	Q4	$\frac{1}{K} \sqrt{\sum_k \frac{\left(x_{ik} - x_{jk} \right)^2}{x_{ik} + x_{jk}}}$	Divergence coefficient, Clark 1952 [1]
Q5	$\frac{\sum_k \left(x_{ik} - \mu_i \right)\left(x_{jk} - \mu_j \right)}{\sqrt{\sum_k \left(x_{ik} - \mu_i \right)^2 \sum_k \left(x_{jk} - \mu_j \right)^2}}$	Correlation coefficient	Q6	$\frac{\sum_k x_{ik} x_{jk} - Km - m\left(\sum_k x_{ik} + \sum_k x_{jk} \right)}{\sqrt{\left(\sum_k x_{ik}^2 - Km^2 - 2m.x_{ik} \right)\left(\sum_k x_{jk}^2 + Km^2 - 2m \sum_k x_{ik} \right)}}$	Cohen 1969 [2]
Q7	$\frac{\sum_k x_{ik} x_{jk}}{\sum_k x_{ik}^2 \sum_k x_{jk}^2}$	Angular distance, Gower 1967 [7]	Q8	$\sum_k^{K-1} \left(\left(x_{ik} - x_{ik+1} \right) - \left(x_{jk} - x_{jk+1} \right) \right)^2$	Meehl Index [10]

The Texture Browsing descriptions had to be transformed from five bins to an eight bin representation in order that all elements of the descriptor measure on an interval scale. A Manhattan metric was used to measure proximity (see [6] for details).

Descriptor extraction was performed using the MPEG-7 reference implementation. In the extraction process each descriptor was applied on the entire content of each media object and the following extraction parameters were used. Colour in Color Structure was quantised to 32 bins. For Dominant Color colour space was set to YCrCb, 5-bit default quantisation was used and the default value for spatial coherency was used. Homogeneous Texture was quantised to 32 components. Scalable Color values were quantised to *sizeof(int)-3* bits and 64 bins were used. Finally, Texture Browsing was used with five components.

These descriptors were applied on three media collections with image content: the Brodatz dataset (112 images, 512x512 pixel), a subset of the Corel dataset (260 images, 460x300 pixel, portrait and landscape) and a dataset with coats-of-arms images (426 images, 200x200 pixel). Figure 1 shows examples from the three collections.

Designing appropriate test sets for a visual evaluation is a highly difficult task (for example, see the TREC video 2002 report [15]). Of course, for identifying the best distance measure for a descriptor, it should be tested on an infinite number of media objects. But this is not the aim of this study. It is just evaluated if – for likely image collections – better proximity measures than those suggested by the MPEG-7 group can be found. Collections of this relatively small size were used in the evaluation, because the applied evaluation methods are above a certain minimum size invariant against collection size and for smaller collections it is easier to define a high-quality ground truth. Still, the average ratio of ground truth size to collection size is at least 1:7. Especially, no collection from the MPEG-7 dataset was used in the evaluation because the evaluations should show, how well the descriptors and the recommended distance measures perform on "unknown" material.

When the descriptor extraction was finished, the resulting XML descriptions were transformed into a data matrix with 798 lines (media objects) and 314 columns (descriptor elements). To be usable with distance measures that do not integrate domain knowledge, the elements of this data matrix were normalised to [0, 1].

For the distance evaluation – next to the normalised data matrix – human similarity judgement is needed. In this work, the ground truth is built of twelve groups of similar images (four for each dataset). Group membership was rated by humans based on semantic criterions. Table 4 summarises the twelve groups and the underlying descriptions. It has to be noticed, that some of these groups (especially 5, 7 and 10) are much harder to find with low-level descriptors than others.

4.2 Evaluation method

Usually, retrieval evaluation is performed based on a ground truth with *recall* and *precision* (see, for example, [3], [16]). In multi-descriptor environments this leads to a problem, because the resulting recall and precision values are strongly influenced by the method used to merge the distance values for one media object. Even though it is nearly impossible to say, how big the influence of a single distance measure was on the resulting recall and precision values, this problem has been almost ignored so far.

In Subsection 2.2 it was stated that the major task of a distance measure is to bring the relevant media objects *as close* to the origin (where the query object lies) *as possible*. Even in a multi-descriptor environment it is then simple to identify the similar objects in a large distance space. Consequently, it was decided to

Table 4. Ground truth information.

Coll.	No	Images	Description
Brodatz	1	19	Regular, chequered patterns
	2	38	Dark white noise
	3	33	Moon-like surfaces
	4	35	Water-like surfaces
Corel	5	73	Humans in nature (difficult)
	6	17	Images with snow (mountains, skiing)
	7	76	Animals in nature (difficult)
	8	27	Large coloured flowers
Arms	9	12	Bavarian communal arms
	10	10	All Bavarian arms (difficult)
	11	18	Dark objects / light unsegmented shield
	12	14	Major charges on blue or red shield

Figure 1. Test datasets. Left: Brodatz dataset, middle: Corel dataset, right: coats-of-arms dataset.

use indicators measuring the distribution in distance space of candidates similar to the query object for this evaluation instead of recall and precision. Identifying clusters of similar objects (based on the given ground truth) is relatively easy, because the resulting distance space for one descriptor and any distance measure is always *one-dimensional*. Clusters are found by searching from the *origin* of distance space to the first similar object, grouping all following similar objects in the cluster, breaking off the cluster with the first un-similar object and so forth.

For the evaluation two indicators were defined. The first measures the average distance of all cluster means to the origin:

$$\mu_d = \frac{\sum_{i}^{no\_clusters} \frac{\sum_{j}^{cluster\_size_i} distance_{ij}}{cluster\_size_i}}{no\_clusters.avg\_distance}$$

where $distance_{ij}$ is the distance value of the j-th element in the i-th cluster, $avg\_distance = \frac{\sum_{i}^{CLUSTERS} \sum_{j}^{cluster\_size_i} distance_{ij}}{\sum_{i}^{CLUSTERS} cluster\_size_i}$, $no\_clusters$ is the number of found clusters and $cluster\_size_i$ is the size of the i-th cluster. The resulting indicator is normalised by the distribution characteristics of the distance measure (*avg_distance*). Additionally, the standard deviation is used. In the evaluation process this measure turned out to produce valuable results and to be relatively robust against parameter p of the quantisation model.

In Subsection 3.2 we noted that p affects the discriminance of a predicate-based distance measure: The smaller p is set the larger are the resulting clusters because the quantisation model is then more discriminant against properties and less elements of the data matrix are used. This causes a side-effect that is measured by the second indicator: more and more un-similar objects come out with *exactly* the same distance value as similar objects (a problem that does not exist for large p's) and become *indiscernible* from similar objects. Consequently, they are (false) cluster members. This phenomenon (conceptually similar to the "false negatives" indicator) was named "cluster pollution" and the indicator measures the average cluster pollution over all clusters:

$$cp = \frac{\sum_{i}^{no\_clusters} \sum_{j}^{cluster\_size_i} no\_doubles_{ij}}{no\_clusters}$$

where $no\_doubles_{ij}$ is the number of indiscernible un-similar objects associated with the j-th element of cluster i.

Remark: Even though there is a certain influence, it could be proven in [5] that no significant correlation exists between parameter p of the quantisation model and cluster pollution.

4.3 Test environment

As pointed out above, to generate the descriptors, the MPEG-7 reference implementation in version 5.6 was used (provided by TU Munich). Image processing was done with Adobe Photoshop and normalisation and all evaluations were done with Perl. The querying process was performed in the following steps: (1) random selection of a ground truth group, (2) random selection of a query object from this group, (3) distance comparison for all other objects in the dataset, (4) clustering of the resulting distance space based on the ground truth and finally, (5) evaluation.

For each combination of dataset and distance measure 250 queries were issued and evaluations were aggregated over all datasets and descriptors. The next section shows the – partially surprising – results.

5. RESULTS

In the results presented below the first indicator from Subsection 4.2 was used to evaluate distance measures. In a first step parameter p had to be set in a way that all measures are *equally* discriminant. Distance measurement is fair if the following condition holds true for any predicate-based measure d_P and any continuous measure d_C:

$$cp(d_P, p) \approx cp(d_C)$$

Then, it is guaranteed that predicate-based measures do not create larger clusters (with a higher number of similar objects) for the price of higher cluster pollution. In more than 1000 test queries the optimum value was found to be $p=1$.

Results are organised as follows: Subsection 5.1 summarises the

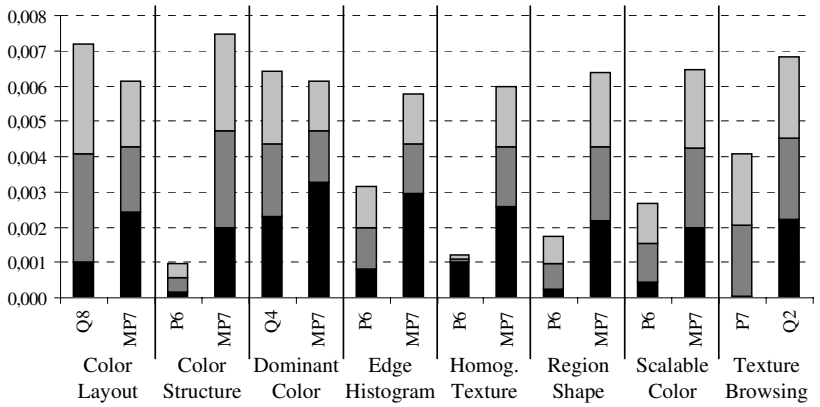

Figure 2. Results per measure and descriptor. The horizontal axis shows the best measure and the performance of the MPEG-7 recommendation for each descriptor. The vertical axis shows the values for the first indicator (smaller value = better cluster structure). Shades have the following meaning: black=μ-σ (good cases), black + dark grey=μ (average) and black + dark grey + light grey=μ+σ (bad).

best distance measures per descriptor, Section 5.2 shows the best overall distance measures and Section 5.3 points out other interesting results (for example, distance measures that work particularly good on specific ground truth groups).

5.1 Best measure per descriptor

Figure 2 shows the evaluation results for the first indicator. For each descriptor the best measure and the performance of the MPEG-7 recommendation are shown. The results are aggregated over the tested datasets.

On first sight, it becomes clear that the MPEG-7 recommendations are mostly relatively good but *never* the best. For Color Layout the difference between MP7 and the best measure, the Meehl index (Q8), is just 4% and the MPEG-7 measure has a smaller standard deviation. The reason why the Meehl index is better may be that this descriptors generates descriptions with elements that have very similar variance. Statistical analysis confirmed that (see [6]).

For Color Structure, Edge Histogram, Homogeneous Texture, Region-based Shape and Scalable Color by far the best measure is pattern difference (P6). Psychological research on human visual perception has revealed that in many situation differences between the query object and a candidate weigh much stronger than common properties. The pattern difference measure implements this insight in the most consequent way. In the author's opinion, the reason why pattern difference performs so extremely well on many descriptors is due to this fact. Additional advantages of pattern difference are that it usually has a very low variance and – because it is a predicate-based measure – its discriminance (and cluster structure) can be tuned with parameter *p*.

The best measure for Dominant Color turned out to be Clark's Divergence coefficient (Q4). This is a similar measure to pattern difference on the continuous domain. The Texture Browsing descriptor is a special problem. In the MPEG-7 standard it is recommended to use it exclusively for browsing. After testing it for retrieval on various distance measures the author supports this opinion. It is very difficult to find a good distance measure for Texture Browsing. The proposed Manhattan metric, for example, performs very bad. The best measure is predicate-based (P7). It works on common properties (*a*, *d*) but produces clusters with very high cluster pollution. For this descriptor the second indicator is up to eight times higher than for predicate-based measures on other descriptors.

5.2 Best overall measures

Figure 3 summarises the results over all descriptors and media collections. The diagram should give an indication on the *general potential* of the investigated distance measures for visual information retrieval.

It can be seen that the best overall measure is a predicate-based one. The top performance of pattern difference (P6) proves that the quantisation model is a reasonable method to extend predicate-based distance measures on the continuous domain. The second best group of measures are the MPEG-7 recommendations, which have a slightly higher mean but a lower standard deviation than pattern difference. The third best measure is the Meehl index (Q8), a measure developed for psychological applications but because of its characteristic properties tailor-made for certain (homogeneous) descriptors.

Minkowski metrics are also among the best measures: the average mean and variance of the Manhattan metric (Q1) and the Euclidean metric (Q2) are in the range of Q8. Of course, these measures do not perform particularly well for any of the descriptors. Remarkably for a predicate-based measure, Tversky's Feature Contrast Model (P1) is also in the group of very good measures (even though it is not among the best) that ends with Q5, the correlation coefficient. The other measures either have a significantly higher mean or a very large standard deviation.

5.3 Other interesting results

Distance measures that perform in average worse than others may in certain situations (e.g. on specific content) still perform better. For Color Layout, for example, Q7 is a very good measure on colour photos. It performs as good as Q8 and has a lower standard deviation. For artificial images the pattern difference and the Hamming distance produce comparable results as well.

If colour information is available in media objects, pattern difference performs well on Dominant Color (just 20% worse Q4) and in case of difficult ground truth (group 5, 7, 10) the Meehl index is as strong as P6.

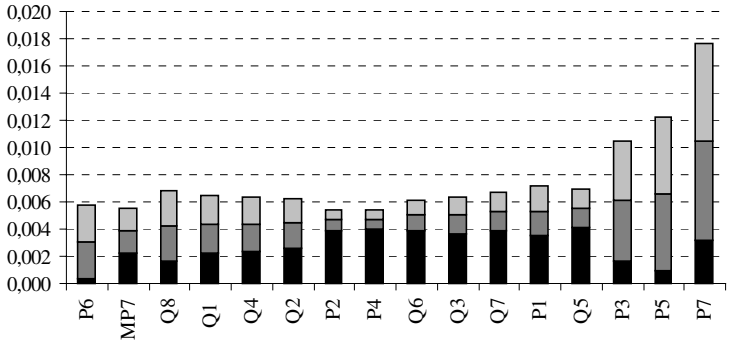

Figure 3. Overall results (ordered by the first indicator). The vertical axis shows the values for the first indicator (smaller value = better cluster structure). Shades have the following meaning: black=μ-σ, black + dark grey=μ and black + dark grey + light grey=μ+σ.

6. CONCLUSION

The evaluation presented in this paper aims at testing the recommended distance measures and finding better ones for the basic visual MPEG-7 descriptors. Eight descriptors were selected, 38 distance measures were implemented, media collections were created and assessed, performance indicators were defined and more than 22500 tests were performed. To be able to use predicate-based distance measures next to quantitative measures a quantisation model was defined that allows the application of predicate-based measures on continuous data.

In the evaluation the best overall distance measures for visual content – as extracted by the visual MPEG-7 descriptors – turned out to be the pattern difference measure and the Meehl index (for homogeneous descriptions). Since these two measures perform significantly better than the MPEG-7 recommendations they should be further tested on large collections of image and video content (e.g. from [15]).

The choice of the right distance function for similarity measurement depends on the descriptor, the queried media collection and the semantic level of the user's idea of similarity. This work offers suitable distance measures for various situations. In consequence, the distance measures identified as the best will be implemented in the open MPEG-7 based visual information retrieval framework VizIR [4].

ACKNOWLEDGEMENTS

The author would like to thank Christian Breiteneder for his valuable comments and suggestions for improvement. The work presented in this paper is part of the VizIR project funded by the Austrian Scientific Research Fund FWF under grant no. P16111.

REFERENCES

[1] Clark, P.S. An extension of the coefficient of divergence for use with multiple characters. Copeia, 2 (1952), 61-64.

[2] Cohen, J. A profile similarity coefficient invariant over variable reflection. Psychological Bulletin, 71 (1969), 281-284.

[3] Del Bimbo, A. Visual information retrieval. Morgan Kaufmann Publishers, San Francisco CA, 1999.

[4] Eidenberger, H., and Breiteneder, C. A framework for visual information retrieval. In Proceedings Visual Information Systems Conference (HSinChu Taiwan, March 2002), LNCS 2314, Springer Verlag, 105-116.

[5] Eidenberger, H., and Breiteneder, C. Visual similarity measurement with the Feature Contrast Model. In Proceedings SPIE Storage and Retrieval for Media Databases Conference (Santa Clara CA, January 2003), SPIE Vol. 5021, 64-76.

[6] Eidenberger, H., How good are the visual MPEG-7 features? In Proceedings SPIE Visual Communications and Image Processing Conference (Lugano Switzerland, July 2003), SPIE Vol. 5150, 476-488.

[7] Gower, J.G. Multivariate analysis and multidimensional geometry. The Statistician, 17 (1967),13-25.

[8] Lance, G.N., and Williams, W.T. Mixed data classificatory programs. Agglomerative Systems Australian Comp. Journal, 9 (1967), 373-380.

[9] Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., and Yamada, A. Color and texture descriptors. In Special Issue on MPEG-7. IEEE Transactions on Circuits and Systems for Video Technology, 11/6 (June 2001), 703-715.

[10] Meehl, P. E. The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In Harlow, L.L., Mulaik, S.A., and Steiger, J.H. (Eds.). What if there were no significance tests? Erlbaum, Mahwah NJ, 393-425.

[11] Pearson, K. On the coefficients of racial likeness. Biometrica, 18 (1926), 105-117.

[12] Santini, S., and Jain, R. Similarity is a geometer. Multimedia Tools and Application, 5/3 (1997), 277-306.

[13] Santini, S., and Jain, R. Similarity measures. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21/9 (September 1999), 871-883.

[14] Sint, P.P. Similarity structures and similarity measures. Austrian Academy of Sciences Press, Vienna Austria, 1975 (in German).

[15] Smeaton, A.F., and Over, P. The TREC-2002 video track report. NIST Special Publication SP 500-251 (March 2003), available from: http://trec.nist.gov/pubs/trec11/papers/VIDEO.OVER.pdf (last visited: 2003-07-29)

[16] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22/12 (December 2000), 1349-1380.

[17] Tversky, A. Features of similarity. Psychological Review, 84/4 (July 1977), 327-351.