Evaluation and Analysis of Similarity Measures for Content-based Visual Information Retrieval

HORST EIDENBERGER

Vienna University of Technology, Institute of Software Technology and Interactive Systems, Interactive Media Systems Group, Favoritenstrasse 9-11, A-1040 Vienna, Austria

Phone 43 1 58801-18853, Fax 43 1 58801-18898

eMail eidenberger@tuwien.ac.at, Web www.ims.tuwien.ac.at

Abstract. The selection of appropriate proximity measures is one of the crucial success factors of content-based visual information retrieval. In this area of research, proximity measures are used to estimate the similarity of media objects by the distance of feature vectors. The research focus of this work is the identification of proximity measures that perform better than the usual choices (e.g. Minkowski metrics). We evaluate a catalogue of 37 measures that are picked from various areas (psychology, sociology, economics, etc.). The evaluation is based on content-based MPEG-7 descriptions of carefully selected media collections. Unfortunately, some proximity measures are only defined on predicates (e.g. most psychological measures). One major contribution of this paper is a model that allows for the application of such measures on continuous feature data. The evaluation results uncover proximity measures that perform better than others on content-based features. Some predicate-based measures clearly outperform the frequently used distance norms. Eventually, the discussion of the evaluation leads to a catalogue of mathematical terms of successful retrieval and browsing measures.

Keywords: *Content-based Visual Information Retrieval, Similarity Measurement, Distance Measurement, Visual Similarity Perception, MPEG-7.*

1. Introduction

The research focus of the experimental study presented in this paper is the identification of proximity measures that can be successfully applied in visual information retrieval (VIR) environments. The term *VIR* refers to a scenario where content-based features are employed to represent the properties of visual media objects and where proximity measures

are utilised to represent human visual similarity perception by distance measurement of feature vectors (vector space model). It is commonly accepted that the careful definition/selection of features and proximity measures are paramount steps to successful VIR. However, so far only relatively few experimental studies on proximity measure selection have been conducted.

This paper offers a number of contributions to the selection problem. We have collected a comprehensive set of similarity measures from various areas of research. These measures are evaluated on content-based features extracted from carefully designed media collections. We have decided on media representation by the visual MPEG-7 descriptors. MPEG-7 is a globally available standard and comprises a well-designed set of heterogeneous features. Unfortunately, some of the employed proximity measures are only defined on predicates (e.g. most psychological and sociological measures). A second major contribution of this paper is a quantisation model that allows for the application of such measures on continuous feature data. Since the quantisation model is defined in a very general context, it can be applied to arbitrarily shaped proximity measures and media descriptions. Eventually, a catalogue of important building blocks of successful similarity measures is derived from the evaluation results. The experiments reveal that particular elements of proximity measures have very specific impacts on the similarity measurement process.

The paper is organised as follows. Section 2 gives background information on similarity measurement and the content-based visual MPEG-7 descriptors. Section 3 introduces the distance measure catalogue. Section 4 sketches the evaluation set-up, including performance indicators and test data. Section 5 discusses the results and gives interpretations. Please note that the following naming conventions are used for mathematical symbols. Vectors (with subscripts, e.g. X_i) and constants (without subscripts, e.g. C) are written in Latin uppercase letters. Latin lowercase letters denote vector elements (e.g. x_{ik}) and variables (e.g. a). Greek lowercase letters are used for weights, thresholds and statistical moments (e.g. μ).

2. Related work

2.1 Similarity measurement for visual data

Generally, similarity measurement on visual information aims at the imitation of human visual similarity perception. Unfortunately, human perception is much more complex than any of the existing similarity models. Human perception includes vision, recognition and subjectivity. The common approach in visual information retrieval is measuring *dis*-similarity as *distance* [17, 13, 8]. Both, query object and candidate object are represented by their corresponding

feature vectors. The distance between the media 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 synthetic/artificial (e.g. properties of a group of objects).

The goal of the measurement process is to express a relationship between the two objects by their distance. Iterating over multiple candidates allows to define a partial order over the media objects and to address those in a (to be defined) neighbourhood as being *similar* to the query object. For the sake of completeness, it has to be mentioned that in a multi-descriptor environment – especially in MPEG-7-based retrieval – this is only half of the way towards a statement on similarity. If multiple descriptors are used (e.g. a description scheme), a rule is required that determines the way how distance values are combined to one global value per media object. However, distance measurement is the most important first step in similarity measurement.

The main challenge for good distance measures is to *reorganise* the description space in a way that media objects with the highest similarity are in close proximity to the query object. If distance is defined minimal (≥ 0), the query object is always in the origin of distance space and similar candidates should form the largest possible clusters near the origin. Most of the frequently used distance measures are based on geometric assumptions of description space (for example, 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). Researchers from different areas have developed alternative models to overcome this shortcoming. Most approaches are predicate-based (descriptors are assumed to contain just binary elements, e.g. Tversky's Feature Contrast Model [35]) and fit better with human perception. We consider distance measures of both groups of approaches in the evaluation.

2.2 The visual MPEG-7 descriptors

The MPEG-7 standard defines – among others – a set of descriptors for visual media. Each descriptor comprises a normative description (in binary and XML format) and guidelines that define how to extract description data and to apply the descriptor to different types of media (e.g. temporal media). The MPEG-7 descriptors have been carefully designed to meet – partially complementary – requirements of different application domains: archival, browsing, retrieval etc. [20, 21]. In this study, we deal exclusively with the content-based visual MPEG-7 descriptors in the context of visual information retrieval and browsing (see [8, 33]).

The visual part of the MPEG-7 standard defines several descriptors [21, 4]. Not all of them are really descriptors in the sense that properties are extracted from visual media. Some of them are just structures for description aggregation and

localisation. The basic descriptors are *Color Layout*, *Color Structure*, *Dominant Color*, *Scalable Color* (colour), *Edge Histogram*, *Homogeneous Texture*, *Texture Browsing* (texture), *Region-based Shape*, *Contour-based Shape* (shape), *Camera Motio*n, *Parametric Motion* and *Motion Activity* (motion) [20, 2].

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 spatio-temporal aggregation and localisation are: *Spatial 2D Coordinates*, *Grid Layout*, *Region Locator* (spatial), *Time Series*, *Temporal Interpolation* (temporal) and *SpatioTemporal Locator* (combined). Supplementary structures exist for colour space representation, colour quantisation and multiple 2D views of 3D objects. These additional structures allow for combining the basic descriptors in a number of ways and on different levels. They do not change the *characteristics* of the extracted information and therefore, aggregation and localisation structures are not considered in the work described in this paper.

As pointed out above, rules are required for the application of features that define how to compute the similarity of two media descriptions. Unfortunately, the MPEG-7 standard does not include distance measures in the normative part, because it was not designed (and should not exclusively be understood) to be retrieval-specific. Nevertheless, the MPEG-7 authors recommend distance measures for their descriptor. These recommendations are based on accurate knowledge of the descriptors' behaviour and the description structures. They are mostly based on L^1 and L^2 metrics (Manhattan distance and Euclidean distance).

3. Distance measures

The distance measures employed in this work have been collected from various areas of research (Subsection 3.1). Because they are defined on differently quantised data ranges, Subsection 3.2 sketches a model for unification on the basis of quantitative descriptions. Subsection 3.3 introduces the distance measures and sketches the original intention of their authors (if known).

3.1 Sources

Distance measurement is used in many research areas including psychology, sociology (e.g. for comparison of test results), medicine (e.g. for comparison of parameters of cases), economics (e.g. for comparison of balance sheet ratios) etc. Naturally, the characteristics of these data differs significantly from area to area. 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. Quantitative values are mostly used for measurements and represent low-level information. Predicates are often employed in psychology, sociology and other human-related sciences. Therefore, most predicate-based distance measures were developed in these areas. Many visual information retrieval descriptions are defined in quantitative terms (as long as semantic enrichment is not involved). Hence, mostly quantitative distance measures are employed in visual information retrieval.

One 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 while some of the most sophisticated distance measures are exclusively defined on predicates, a model is required that allows for the application of predicate-based distance measures on quantitative data. Such a model – developed for this study – is introduced in the next section.

3.2 Quantisation model

The purpose of the quantisation model is to extend the set operators that are employed in predicate-based distance measures to the continuous domain. The first in visual information retrieval to follow this approach were Santini and Jain, who tried to apply Tversky's Feature Contrast Model [35] to content-based image retrieval [30, 31]. They interpreted continuous data as fuzzy predicates and made use of fuzzy set operators. Unfortunately, their model suffered from several shortcomings they described in [30, 31]. For example, the quantitative model worked only for one specific version of the original Feature Contrast Model.

The main idea of the presented quantisation model is that set operators are replaced by *statistical* functions. In [10] the authors show that this interpretation of set operators is a reasonable approach. 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, we show that the model does not only work for predicates but for quantitative data as well. Each measure that implements the model can be used as a substitute for the original predicate-based measure.

Generally, the binary properties of two objects (e.g. media objects) may exist in both objects (the sum of these properties is 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 described in equation 1 (see [10] for further details). M, μ, σ are span, mean and variance of the elements of the data vectors X_i. By convention, $x_{max}=1$ and $x_{min}=0$.

(Please refer to the last paragraph of Section 1 for the naming conventions of mathematical symbols used in this paper.)

$$
\begin{aligned}
a &=< X_i \cup X_j, 1 >= \sum_k s_k, \quad s_k = \begin{cases} \frac{x_{ik} + x_{jk}}{2} & if\ M - \frac{x_{ik} + x_{jk}}{2} \le \varepsilon_1 \\ 0 & otherwise \end{cases} \\
b &=< X_i - X_j, 1 >= \sum_k s_k, \quad s_k = \begin{cases} x_{ik} - x_{jk} & if\ M - (x_{ik} - x_{jk}) \le \varepsilon_2 \\ 0 & otherwise \end{cases} \\
c &=< X_j - X_i, 1 >= \sum_k s_k, \quad s_k = \begin{cases} x_{jk} - x_{ik} & if\ M - (x_{jk} - x_{ik}) \le \varepsilon_2 \\ 0 & otherwise \end{cases} \\
d &=< \neg X_i \cup \neg X_j, 1 >= \sum_k s_k, \quad s_k = \begin{cases} M - \frac{x_{ik} + x_{jk}}{2} & if\ \frac{x_{ik} + x_{jk}}{2} \le \varepsilon_1 \\ 0 & otherwise \end{cases}
\end{aligned}
\tag{1}
$$

with:

$$
\begin{aligned}
& X_i = (x_{ik})\ with\ x_{ik} \in [x_{\min}, x_{\max}] \\
& M = x_{\max} - x_{\min} \\
& \varepsilon_1 = \begin{cases} M\left(1 - \frac{\mu}{f}\right) & if\ f \ge \mu \\ 0 & else \end{cases} \quad where\ \mu = \frac{\sum_i \sum_k x_{ik}}{i * k} \\
& \varepsilon_2 = \begin{cases} M\left(1 - \frac{\sigma}{f}\right) & if\ f \ge \sigma \\ 0 & else \end{cases} \quad where\ \sigma = \sqrt{\frac{\sum_i \sum_k (\mu - x_{ik})^2}{i * k}} \\
& f \in \Re^+ \setminus \{0\}
\end{aligned}
$$

Term *a* stands for 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* sums those properties that are absent in both data vectors. Every property is selected/weighted 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 of all elements of description space).

The implementation of these operators is based on the assumption that all feature vector elements measure on interval scale. In simple words, each element has to express a property that is "more or less" present ("*0*": not at all, "*M*": fully present). This is true for most visual descriptors and all considered MPEG-7 descriptors. A natural origin, as it is assumed here ("*0*"), is not required.

The quantisation model is solely controlled through parameter *f* (called discriminance-defining factor). Parameter *f* is an additional criterion for the behaviour of a distance measure and determines the thresholds used by the operators (ε_1, ε_2).

It expresses the accuracy of data items (quantisation) and therefore, how accurate they should be investigated. Parameter *f* can be set by the user or automatically. Interesting are the limits:

$$f \to \infty \Rightarrow \varepsilon_1, \varepsilon_2 \to M \tag{2}$$

In this case, all elements (properties) are assumed to be continuous (high quantisation). All properties of a description are used by the operators. The distance measure is *not* discriminative for properties.

$$f \to 0 \Rightarrow \varepsilon_1, \varepsilon_2 \to 0 \tag{3}$$

Now, all properties are assumed to be predicates. Only binary elements (predicates) are used by the operators (1-bit quantisation). The distance measure is highly discriminative for properties.

Between these limits, every distance measure that uses the quantisation model is – depending on *f* – more or less discriminative for properties. That is, 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 exclusively of binary elements, *f* should be used as follows (for example, *f* can easily be set automatically):

$$f \to 0 \Rightarrow \varepsilon_1, \varepsilon_2 = 0, e.g.\ f = min(\mu, \sigma) \tag{4}$$

In this case, the measurements of *a*, *b*, *c*, *d* have the same characteristics as 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. In fact, the quantisation model works accurately on predicate data for any $f \neq \infty$.

In order to illustrate 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 distance) degenerates to the Hamming distance (from [20], without weights):

$$L^1 = \sum_k \left| x_{ik} - x_{jk} \right| \equiv b + c = Hamming\ distance \tag{5}$$

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, *f* should be used as follows (again, *f* can easily be set automatically):

$$f \to \infty \Rightarrow \varepsilon_1, \varepsilon_2 = 1 \tag{6}$$

In this situation, when *f* approaches infinity, *a* and *d* become continuous functions:

$$\begin{aligned} M - \frac{x_{ik} + x_{jk}}{2} \le M \equiv true \Rightarrow a = \sum_k s_k \; where \; s_k = \frac{x_{ik} + x_{jk}}{2} \\ \frac{x_{ik} + x_{jk}}{2} \le M \equiv true \Rightarrow d = \sum_k s_k \; where \; s_k = M - \frac{x_{ik} + x_{jk}}{2} \end{aligned} \tag{7}$$

With *f* approaching infinity, *b* and *c* are continuous for the following expressions:

$$\begin{aligned} & M - \left(x_{ik} - x_{jk}\right) \le M \equiv x_{ik} - x_{jk} \ge 0 \\ & \Rightarrow b = \sum_k s_k \; where \; s_k = \begin{cases} x_{ik} - x_{jk} & if \; x_{ik} - x_{jk} \ge 0 \\ 0 & else \end{cases} \\ & M - \left(x_{jk} - x_{ik}\right) \le M \equiv x_{jk} - x_{ik} \ge 0 \\ & \Rightarrow c = \sum_k s_k \; where \; s_k = \begin{cases} x_{jk} - x_{ik} & if \; x_{jk} - x_{ik} \ge 0 \\ 0 & else \end{cases} \\ & \Rightarrow b + c = \sum_k s_k \; where \; s_k = \left| x_{ik} - x_{jk} \right| \\ & \quad b - c = \sum_k s_k \; where \; s_k = x_{ik} - x_{jk} \\ & \quad c - b = \sum_k s_k \; where \; s_k = x_{jk} - x_{ik} \end{aligned} \tag{8}$$

This means, for a sufficiently high value of *f* every predicate-based distance measure that either abandons *b* and *c* or uses just the terms *b*+*c*, *b*-*c* or *c*-*b*, can be transformed into a continuous quantitative distance measure. For example, the Hamming distance (again, without weights):

$$b + c = \sum_k s_k \; where \;\; s_k = \left| x_{ik} - x_{jk} \right| = \sum_k \left| x_{ik} - x_{jk} \right| = L^1 \tag{9}$$

The quantisation model successfully reconstructs the L^1 metric without any distance measure-specific modifications to the model. This demonstrates the reasonability of the quantisation model. In the following evaluation it will be employed to extend successful predicate-based distance measures on the quantitative domain.

In summary, 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 (4) parameter *f* allows for controlling the similarity measurement process in a novel way (by discriminance on property level). However, the application of the quantisation model requires the identification of appropriate values for *f*. Apart from the suggestions above, it is not possible to give general rules for the optimal selection of *f*. The sensitivity of the parameter depends heavily on the characteristics of the feature data (distribution etc.). Fortunately, a simple algorithm solves the problem:

```
1. Select sample vectors from the feature data set
```

2. Initialise parameter *f=1*
3. For each pair of vectors from the sample: compute L^1 metric, Hamming distance
4. Aggregate the pair-wise distance values to two average distances
5. If the average distances do not match: adapt *f* and return to 3.

We make use of the equivalency of the Manhattan metric and the Hamming distance to identify *f* values that guarantee that the predicate-based distance measures are equally discriminative as the quantitative distance measures. The adaptation in step 5 should be performed as discussed above for the limits of *f*.

3.3 Implemented measures

For the evaluation described in this work we implemented predicate-based (based on the quantisation model), quantitative and the distance measures recommended in the MPEG-7 standard. In total, 37 different distance measures were evaluated.

Table 2 summarises those predicate-based measures that performed best in the evaluation. Twenty predicate-based measures were investigated. For all 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 [35]) and α, β are weights for elements b and c. In [10] 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-P6) but have been heavily exploited in psychological research. Goodall, for instance, investigated the behaviour of the simple match coefficient (P6) for independent predicates [14]. P9 was very successful in measurement of the similarity of multi-level variables. Sokal and Sneath and others developed several measures that stress co-presence of predicates (P7, P10-P12, P14, P16, P17) [17]. P13 is a further-developed version of P8. Pattern difference (P18) is used in the statistics package SPSS for cluster analysis. P20 is a correlation coefficient for predicates developed by Pearson [27]. P18, P19, P20 are similar measures, since all of them make use of the term $b*c$. In our evaluations, the product of differences turned out to be a very powerful model for human similarity perception.

Table 3 lists the quantitative distance measures that were investigated. Q1 and Q2 are metrics-based and were implemented as representatives for the entire group of Minkowski distances. Q3, the Canberra metric, is a normalised version of Q1. Similarly, Q4, Clark's divergence coefficient is a normalised version of Q2. In Q5, μ_i is the mean of the elements of description X_i. In Q6, m is $\frac{M}{2}$ (=0.5). 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. In Q7, μ, σ are mean and standard deviation over all elements of X_i and X_j. In Q8, d_{ij} is the Euclidean distance as defined in Q2 (without weights). Q8 was developed by Catell to compare psychological profiles. Finally, Q10 is a measure that takes the differences between adjacent vector elements into account. This property 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 employed, 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). For 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 function all $\alpha(t)$ were set to "1" and matching was performed rotation- and scale-invariant. It is important to notice that some of the measures presented in this section are *distance* measures while others are *similarity* measures. For the purpose of the evaluation all similarity measures were transformed to distance measures.

4. Evaluation set-up

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

4.1 Performance indicators

Usually, retrieval and browsing evaluation are based on a ground truth and *recall* and *precision* indicators (see, for example, [8, 33]). In multi-descriptor environments this approach leads to a problem, since the recall and precision values are strongly biased by the method used to merge the distance values of media objects. Though it is nearly impossible to estimate the influence of a single distance measure on the final recall and precision values, the merging problem has been frequently ignored so far. In Subsection 2.1 it was stated that the major task of a distance measure is to put 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. Hence, we decided to define

performance indicators that measure the *distribution* of objects in distance space instead of recall- and precision-like measures.

Our performance indicator should measure two properties. Firstly, it should take the size of positive clusters into account (larger clusters are better). Positive clusters are clusters of objects similar to the query example. That is, they belong to the same ground truth group. We call this property *browsing* property, since for browsing applications it is important that similar objects are grouped in close proximity to each other. In fact, this measure may be best described as a *browsing measure for retrieval results*. It quantises the browsing qualities of retrieval results.

Secondly, the performance indicator should consider the distance of similar objects to the origin of distance space. The query example lies in the origin, because its distance to itself is always zero. Similar objects should be positioned as close to the query example as possible. This property is called *retrieval* property, because for retrieval applications it is important that as many similar objects as possible are among the first results.

The browsing property could be expressed as average cluster size by the following term, in which C is the number of clusters and c_i is the size (in elements) of the i-th cluster.

$$p^{BROWSING*} = \frac{\sum_{i=1}^{C} c_i}{C} \tag{10}$$

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 object that belongs to the same group as the query example, grouping all following similar objects (same ground truth group) in the cluster, breaking off the cluster with the first un-similar object (different ground truth group) and so forth. See Figure 2 for an example with three clusters of similar objects. Equation 11 expresses the retrieval property as the average distance of objects similar to the query example. Here, d_{ij} is the distance of the j-th element of the i-th cluster. The other symbols are defined as above.

$$p^{RETRIEVAL*} = \frac{\sum_{i=1}^{C} \sum_{j=1}^{c_i} d_{ij}}{\sum_{i=1}^{C} c_i} \tag{11}$$

$p^{BROWSING*}$ measures on [1, S], where S is the number of similar media objects in the evaluated collection (denominator of equation 11). Optimally, distance measures should maximise $p^{BROWSING*}$. For distance measures defined on [0, 1],

$p^{RETRIEVAL*}$ measures on [0, 1]. The best measures should minimise the average distance. Since we want to use both measures in combination, we redefine $p^{BROWSING*}$ in the following way (using inversion, subtraction of *1/S* and normalisation by *(S-1)/S*). $p^{BROWSING}$ measures on [0, 1]. The best values are near zero.

$$p^{BROWSING} = \frac{C-1}{\sum_{i=1}^{C} c_i - 1} \tag{12}$$

Note that $p^{BROWSING}$ is only defined for collections with at least two elements! Furthermore, in the presented form, $p^{RETRIEVAL*}$ can only be employed to compare distance measures that measure on the same interval. Since this is not the case for most of the measures used in this evaluation, we redefine $p^{RETRIEVAL*}$ in the following way.

$$p^{RETRIEVAL} = \frac{\sum_{i=1}^{C}\sum_{j=1}^{c_i} d_{ij}^{'}}{\sum_{i=1}^{C} c_i} \quad with \quad d_{ij}^{'} = \frac{d_{ij} - d_{\min}}{d_{\max} - d_{\min}} \tag{13}$$

All distance values are transformed to [0, 1] using the minimum distance value d_{min} and the maximum d_{max}.

Eventually, we define the weighted sum of the two basic properties as the performance indicator (equation 14). It is important to notice that the browsing indicator and the retrieval indicator are interdependent measures, i.e. a good distance measure should optimise both measures. If just the browsing indicator is optimised by a distance measure it may not be able to identify relevant media objects as the best matches. Distance measures that optimise only the retrieval indicator may fail in distinguishing similar from unsimilar media objects.

$$p = \alpha\, p^{RETRIEVAL} + \beta\, p^{BROWSING} \quad with \quad \alpha, \beta \geq 0 \wedge \alpha + \beta = 1 \tag{14}$$

The value of *p* is independent of collection size and distance measure used. For the evaluations presented below, we use weights of 0,5 for α, β. We investigate mean and standard deviation of *p*, $p^{BROWSING}$ and $p^{RETRIEVAL}$ over series of test queries. It has to be noted that the combination of $p^{BROWSING}$ and $p^{RETRIEVAL}$ values would be questionable if these variables had different variances. However, as we found in the evaluation process, both measures come up with highly similar variances (over a sufficiently large number of tests). Therefore, overall performance results are not biased by the linear combination. In the evaluation process these measures turned out to provide valuable results and to be robust against parameter *f* of the quantisation model.

4.2 Test data

For the evaluation seven MPEG-7 descriptors were used. All colour descriptors: *Color Layout*, *Color Structure*,

Dominant Color, *Scalable Color*, two texture descriptors: *Edge Histogram*, *Homogeneous Texture* and one shape descriptor: *Region-based Shape*. *Texture Browsing* was not employed, because the MPEG-7 standard suggests that it is not suitable for retrieval. The other basic shape descriptor, *Contour-based Shape*, was not considered, because it produces structurally different descriptions that cannot be transformed to data vectors with elements measuring on interval scales. The motion descriptors were not employed, 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. Finally, no high-level descriptors were considered (*Localisation*, *Face Recognition* etc., see Subsection 2.2), 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.

Description extraction was performed using the MPEG-7 eXperimentation Model [23]. 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. The *Dominant Color* colour space was set to YCrCb, 5-bit default quantisation was employed and the default spatial coherency algorithm was applied. *Homogeneous Texture* was quantised to 32 components. Finally, *Scalable Color* values were quantised to *sizeof(int)-3* bits and 64 bins.

The descriptors were applied on several media collections with varying content (image libraries and video clips). For workflow optimisation we implemented a web interface for submission and evaluation of media descriptions. For the evaluations presented below, we selected three media collections with image content from the evaluated datasets: the Brodatz dataset (112 greyscale images of textures, 512x512 pixel), a subset of the Corel dataset (260 colour photos of humans, animals and flowers, 460x300 pixel, portrait and landscape) and a dataset with coats-of-arms images (426 synthetic colour images, 200x200 pixel). Figure 1 depicts examples from the three collections. We used collections of this relatively small size, because the applied evaluation methods are invariant for collection size above a certain minimum size. Additionally, it is easier to define a high-quality ground truth for smaller collections. Still, the average ratio of ground truth size to collection size is at least 1:7. Coats-of-arms, Brodatz and Corel were chosen, because – to our experience – they reveal the characteristic properties of distance measures (see, for example, results in [9, 12]). The same characteristic properties were also identified using other collections, but with less striking performance indicator values. These results may origin from the somewhat less differentiated visual content. For example, the evaluated video clips (news, advertisements, documentaries etc.) consist of groups of very similar frames. On the other hand, inter-group similarity hardly exists. In consequence, evaluation results have lower variance.

Generally, designing appropriate test sets for visual evaluation is a highly sophisticated task (for example, see [26, 1]).

Almost every selection of media objects can be argued against. We are aware that for our goal, identifying the best distance measure for descriptors, the distance measures should be tested on a large number of media objects. In order to support the reader in evaluating his own media descriptions, we decided to open the evaluation website – described in the next subsection – to the public. It can be accessed from [24] and readers are invited to use it.

For the distance evaluation – next to media descriptions – human similarity judgement is needed. For the coats-of-arms, Brodatz and Corel datasets we defined twelve groups of similar images (four for each dataset). Group membership was selected by human testers based on semantic criteria. Table 4 summarises the twelve groups and the underlying descriptions. It has to be noted, that some of these groups (especially 5, 7 and 10) are much harder to identify by low-level descriptions than others, because they are defined on a semantically higher level.

4.3 Test environment

The distance measure evaluation framework was implemented with a website front-end. Figure 3 illustrates the workflow for submission and evaluation. After submission, the entered data are inspected by the website administrator: If the data do not contain illegal content (e.g. copyright-protected media objects), evaluation is performed in a background process. Evaluation results are published on a second webpage.

The submission procedure parses the description elements from the XML descriptions and transforms them into a data matrix with one line per media object and one column per description element (e.g. 318 columns for all seven evaluated MPEG-7 descriptors). The elements of this data matrix have to be normalised in order to be usable with general-purpose distance measures. We perform a simple column-wise min-max-normalisation.

$$x'_{ij} = \frac{x_{ij} - min_j}{max_j - min_j} \tag{15}$$

min_j is the minimum and max_j is the maximum of column j. The resulting values x'_{ij} are normalised to [0, 1]. The distribution of elements is not affected by this operation. The data matrix and the rest of the submission data (including ground truth information) are stored in a database.

The evaluation procedure runs 100 queries for each distance measure. Each iteration is a sequence of the following steps: (1) random selection of a ground truth group, (2) random selection of a query object from this group, (3) distance measurement to all other objects in the dataset, (4) clustering of the resulting distance space based on the ground truth and finally, (5) computation of performance indicators. Mean and standard deviation of p, $p^{BROWSING}$ and $p^{RETRIEVAL}$ for each descriptor are published on the results webpage. Data tables for the coats-of-arms, Brodatz and Corel datasets can

be found in the web appendix [36].

5. Results

In the results presented below, the performance indicators from Subsection 4.1 are used to evaluate distance measures. Weights α, β of performance indicator p (equation 14) are set to 0,5 each. Parameter f of the quantisation model is set to $f=1$, because – as we found in experiments – in this case predicate-based and quantitative distance measures are approximately equally discriminative (please refer to Section 3.2 for details). The results section is organised as follows. Subsection 5.1 compares the best-performing distance measures to the MPEG-7 recommendations and Subsection 5.2 analyses the terms frequently occurring in successful distance measures. Please note that the results presented below are based on the evaluation results that can be found in the web appendix [36].

5.1 Best distance measures for MPEG-7 descriptions

Table 5 summarises the performance of the best measures in comparison to the distance measures recommended by the MPEG-7 group. The results are averaged over the coats-of-arms, Brodatz and Corel datasets. Tests on other collections confirm these results. As can be seen, the best measures are always predicate-based. This fact supports our argumentation in Subsection 3.2 that the quantisation model represents a reasonable approach. Overall, for most descriptors the p values of the best measures are between 60% and 70% of the respective MPEG-7 measures (smaller is better). Exceptions are *Edge Histogram* and *Homogeneous Texture*, where P19 (Yule coefficient) and P18 (pattern difference measure) perform substantially better. Looking at $p^{RETRIEVAL}$ (retrieval quality) and $p^{BROWSING}$ (browsing quality) we can see that the major difference is in the retrieval quality. The best distance measures put similar media objects significantly closer to the origin of distance space than the MPEG-7 recommendations. The best $p^{BROWSING}$ values are not bad, but significantly higher than the $p^{RETRIEVAL}$ values.

In Figures 4-6 we have a closer look at the best- and worst-performing distance measures. The figures are summarised over the three considered media collections and all seven descriptors. Diagrams a, b, c depict the numbers of appearances among the top three, top ten and worst ten distance measures for the p, $p^{RETRIEVAL}$, $p^{BROWSING}$ values, respectively. Several observations can be made. P1, P19, P20 (Pearson coefficient) seem to be the best measures in terms of retrieval quality. For all descriptors (except *Color Layout*, as will be shown below) these measures outperform the other measures and the MPEG-7 recommendations. Since all of these measures make use of description elements that are present in both media objects (a property), a seems to be important for minimising the distance of similar media

objects in distance space. On the contrary, P18 (pattern difference) optimises the browsing quality. This is not unexpected as P18 is widely used in cluster analysis algorithms. Data clustering and visual media browsing follow the same optimisation criterion. Interestingly, P18 does not make use of *a*, the most prominent term is *b***c*. It seems that *a* is counterproductive for maximising cluster sizes. Together, P18, P19 and P20 are the most robust distance measures that outperform most other measures in both retrieval and browsing scenarios. All three utilise the term *b***c* (product of differences in two media objects) that appears superior in many situations over alternative formulations of the same aspect (especially, *b*+*c*, e.g. used in P1). Below, we have a closer look on this phenomenon. Parameter *K* (number of considered predicates, $K=a+b+c+d$), even though being variant (determined by parameter *f* of the quantisation model), does not seem to have an influence on retrieval quality or browsing quality.

Only Q7 (Webster's intra-class coefficient) and Q8 (Catell's measure) of the quantitative distance measures are able to compete with the predicate-based measures. Both perform well in retrieval and browsing scenarios. Q8 has an especially high browsing performance for the *Dominant Color* and the *Edge Histogram* descriptors. On *Color Layout* and *Homogeneous Texture* descriptions it performs poorly. Since the elements of the first two descriptors generally have a high variance (as we found in [12]) while *Color Layout* and *Homogeneous Texture* elements have a very low variance, it seems that Q8 is sensitive to data variance. Indeed, the main component of the measure is a square-rooted Euclidean distance that measures distance as the average difference of data values. Another interesting point is that Q7 and Q8 perform unsatisfactorily on colour descriptions of monochrome content (Brodatz dataset). Since these data vectors contain mostly zero values (and, in consequence, have low variances), quantitative measures are unable to fulfill retrieval and browsing tasks properly. Besides, the sensitivity of Q7 and Q8 explains, why both measures appear frequently among the top three measures (for some descriptors) and among the worst ten measures (for other, less suitable descriptors). The most stable quantitative measure is Q10 that measures differences of differences in neighbouring data values. Q10 is hardly ever among the best distance measures but performs average in retrieval and browsing scenarios. P9 is the only predicate-based measure that performs as bad as the quantitative distance measure. Combining *a*+*d*, *b*+*c* and *K* as in P9 results in a measure with similar properties as the quantitative distance measures have.

Figures 7-13 depict the best five distance measures per descriptor and the MPEG-7 recommendations. For the three performance indicators, mean and standard deviation (averaged over the three considered media collections) are shown. These diagrams allow for more detailed analysis of the best distance measures. For *Color Layout*, *Color Structure* and *Region-based Shape* P19 performs best. P19 is especially good in optimising the retrieval quality. P6 is the best measure

for *Scalable Color* and second best for *Color Layout*. As pointed out above, these descriptors have low variance and contain many zero values. In consequence, P6 is (despite high variance) a suitable measure. P18 performs best on *Homogeneous Texture* and well on most other descriptors. This is due to its outstanding browsing performance. P1 (Feature Contrast Model) performs best on *Dominant Color* and *Edge Histogram*. The major strength of the P1 measure is outstanding retrieval performance (similar to P19). It is interesting to notice that the MPEG-7 distance measures, even though performing worse, mostly have a lower standard deviation than the best measures. The reason may be the integration of domain knowledge in the MPEG-7 distance measures (e.g. weights in *Color Layout* descriptor) that make them more robust against variances in the description elements.

Looking at distance terms used in the measures, we can confirm that $b*c$ is the most important term for browsing quality. Especially if the variance in the data elements is low, building the product of differences helps distinguishing clusters of similar objects from the remaining objects. Psychological research on human visual perception has revealed that in many situations differences between the query object and a candidate weigh much stronger than common properties. The pattern difference measure, that exclusively relies on $b*c$, implements this aspect in the most consequent manner. In comparison, $b+c$ is a valuable term for both retrieval and browsing that performs well independently of the investigated media collection type (low standard deviation for performance indicators). But, in contrast to $b*c$, $b+c$ reaches its highest performance only for description elements with high variance. For others, e.g. *Homogeneous Texture*, it falls behind $b*c$-based measures (especially, in browsing performance). The term $a+d$ performs well for browsing purposes (independently of variance in the data). The term $a*d$ performs poorly. Clusters are small and retrieval performance is only average. It seems that for properties that are present/not present in both media objects, using the product of a and d gives too much power to d. In the next subsection we will see that d can play an important role to improve distance measurement performance. However, employing it to weight a may go too far.

In conclusion of this subsection, selected predicate-based measures (based on the quantisation model) clearly outperform the MPEG-7 recommendations. Remarkably, even though the MPEG-7-recommended distance measures make use of domain knowledge and the meaning of description elements, general-purpose measures exist that are able to compute distance spaces of higher retrieval and browsing quality. For example, the Feature Contrast Model performs better on *Dominant Color* descriptions than the tailor-made MPEG-7 distance measure. It would be a promising piece of future work to further optimise the best distance measures by integrating the same MPEG-7 domain knowledge. Additionally, we identified patterns in the best measures that appear to be important success factors. In the next subsection we will have a closer look on these "ingredients" of successful distance measures.

5.2 Analysis of successful distance measures

In Subsection 5.1 we saw that *b*c* (P18) appears to be the best browsing term, while using *b+c* (P1, P12) leads to better retrieval results than *b*c* alone. It would be interesting to identify the variation of b+c that performs best: *b* (as in P1, weight β is zero!) or *b+c* (as in P12, P3 etc.). Investigating the data tables (please see the web appendix [36]) we find that in about 75% of all cases that *b* outperforms *b+c*. The major difference is in retrieval quality. The browsing quality is about the same. The superiority of *b* and earlier findings in [10] justify eliminating the β*c term in P1.

In browsing scenarios we find that *a+d* (P6, P12 etc.) and *a* (P1) are patterns that perform well. It would be interesting to know, under which circumstances it makes sense to employ *a* alone and when to use it in combination with *d* (*d* stands for properties present in neither of the compared media objects). From closer analysis we can see that *a+d* is superior over *a* in about 66% of all cases. This term performs sometimes better than others in terms of browsing quality. If *d* is helpful, it usually improves performance to excellent values. Additionally, using *a+d* performs better than *a* if the investigated data contain many zero values (e.g. *Scalable Color* descriptor, Figure 13). Then, it makes the distance measure more sensitive if absent properties are explicitly included.

Distance measurement for sparsely populated descriptions is a problem of general interest. We would like to see whether terms exist (next to *a+d*) that perform better than average. Analysing the Brodatz dataset we can see that the best measures are P6 and P4. Structurally similar measures, like P7 and P10, perform better as well. However, frequently P6 outperforms P4, because of the usage of *K* (number of considered predicates/description elements). Elements with zero variance are never considered in the quantisation model and reduce the value of *K*. P6 is a similarity measure. For distance measurement we use the inverse form. Then, a lower *K* reduces the distance value and therefore, has a discriminating effect. Interestingly, P1 performs relatively poor on sparsely populated descriptions. The terms *a* and *b* alone may not have enough variance to distinguish similar objects successfully. The same is true for Q8. We already saw that quantitative measures depend on high variance in the descriptions. P18, P19, P20 (using *b*c*) perform as well as for content with higher variance.

Summarising our findings so far, we attempt building a hierarchy of successful distance measure patterns. Since most terms have a particular strength, we try to order patterns according to their retrieval and browsing performance. Of course, this order cannot be absolutely correct for any type of media and descriptor. We do not claim it being generally true. It should only help to distil and illustrate our experiences. Below, an expression "*A* >> *B*" means, that if pattern *A* is used in a distance measure, then this measure is in average superior over a second measure using pattern *B*. We find the following hierarchies being appropriate:

Retrieval: $a*d-b*c >> a-b >> a*d >> b+c >> b*c >> a+d >> a$ (16)

Browsing: $b*c >> a+d >> b >> b+c >> a*d$ (17)

In equations 16, 17, similarity and distance terms are mixed. Parameter *K* is not considered, as it is only relevant for descriptions with many zero elements. It can be seen that using *a* (common properties) is of highest importance for retrieval purposes. For browsing, using *b* (properties that appear in the query example) is crucial. One could argue that an optimal distance measure could simply be created by employing the best patterns from above and combining them appropriately. Some measures show the opposite. P9, for example, combines $a+d$ and $b+c$, but appears among the worst performing distance measures. Instead, (predicate-based) distance measures should still be derived from qualitative considerations on the tackled problem domain.

Furthermore, it would be interesting to identify why some patterns perform better than others. Specifically, we compare $b*c$ to $b+c$ and $a*d$ to $a+d$. In the quantisation model (defined in equation set 1), *b*, *c* are defined as the sums of non-negative differences $x_{ik}-x_{jk}$ over some description elements. The conditions are not relevant here, since they are the same for $b*c$ and $b+c$. The span of x_{ik} is *M*. Therefore, $b*c$ is defined on $[0, k*M^2]$ while $b+c$ is defined on $[0, 2.k*M]$ only. Variable *k* is the number of considered elements ($k \leq K$). In consequence, $b*c$ is more discriminative, i.e. it has more *power* for distributing data vectors with low variance on a wider range. Since the operators of the quantisation model take care that only relevant description elements are selected, it is more likely that similar elements (low *b*, *c*) are positioned close to each other. The same argumentation is true for $a*d$ and $a+d$. The major difference is that *a* and *d* measure properties existing in both/none of the descriptions. Therefore, $a*d$ is a better retrieval term than $a+d$ (bringing descriptions close to the origin of distance space) while $b*c$ is a better browsing term than $b+c$ (bringing similar descriptions close to each other).

In conclusion of this subsection, Table 6 summarises some noteworthy results. For the argumentation we make use of the hierarchies defined in equations 16, 17. On *Edge Histogram* descriptions (high variance, see [12]) P1 outperforms P3 and P8 in retrieval quality. P3 does not employ *a*, P8 uses *c* in addition. As we found above, the best combination is to use *a* and *b*. On the same data, Q8 performs better than Q2 (Euclidean distance), even though it uses the Euclidean distance. The difference of these two measures is that Q8 uses *K* and reduces the influence of the Euclidean distance by taking the squared root. On *Homogeneous Texture* descriptions (low variance) P20 outperforms P16. Since the only difference between P16 and P20 is the usage of $b*c$, this finding illustrates the positive influence of this term. For *Scalable Color* descriptions (very low variance) it can be observed that P18 performs much better than P1 (especially, in terms of browsing quality). In our opinion this is due to the usage of the $b*c$ pattern. Since $b*c$ is worse than $b+c$ in

retrieval quality, P18 performs worse than P1 for $p^{RETRIEVAL}$.

5.3 Summary

Below, the major findings of this section are summarised.

- Type of descriptors used, characteristics of the queried media set and the properties of human similarity perception determine the selection of suitable distance measures.
- Proximity measures can be identified that outperform the distance measure recommendations of the content-based visual descriptors proposed in the MPEG-7 standard. The best predicate-based and quantitative distance measures achieve 30%-40% better results than the MPEG-7 measures.
- The predicate-based distance measures perform in average better than the quantitative measures. In particular, the measures P18 (pattern difference), P19 (Yule coefficient) and P20 (Pearson coefficient) achieve the highest performance. Retrieval quality is maximised by measures Tversky's Feature Contrast Model P1, the Yule coefficient P19 and the Pearson coefficient P20. The pattern difference measure P18 delivers the highest browsing quality.
- Of the quantitative measures, only Q7 (Webster's intra-class coefficient) and Q8 (Catell's measure) achieve results comparable to the best predicate-based measures. In average, Meehl's index Q10 is the quantitative measure that produces the most reliable results. It can be observed that the performance of the quantitative measures depends on the existence of high variance in the descriptions.
- Equations 16 and 17 establish a rough order of successful distance expressions. The term $b*c$ is the paramount term for browsing quality. The term $b+c$ maximises the retrieval quality.

6. Conclusions

The study presented in this paper evaluates a set of distance measures for their suitability for feature-based visual information retrieval. Additionally, we suggest a model for the unification of predicate-based and continuous distance measures and derive successful distance patterns from the evaluation results. There are 37 proximity measures utilised on seven MPEG-7 descriptors and three media collections. Performance indicators are defined and more than 75000 tests are performed.

In the evaluation the best overall distance measures for visual content – as extracted by the visual MPEG-7 descriptors –

turn out to be the pattern difference measure, the Feature Contrast Model, the Pearson and the Yule coefficient. Since these four measures perform significantly better than the MPEG-7 recommendations, we recommend to examine them in more detail (e.g. on media collections from [26]) and to use them for content-based retrieval and browsing applications.

The major performance gap of proximity measures is often in the retrieval quality. In browsing-like scenarios, the best measures are usually only slightly better than the distance measures recommended in the MPEG-7 standard. Generally, we found that predicate-based distance measures perform significantly better than quantitative measures (e.g. Euclidean distance). This finding supports the introduction of the quantisation model. The quantisation model opens an entirely new range of possibilities for similarity measurement in content-based visual information retrieval and browsing.

In summary, the choice of the most suitable distance function for similarity measurement depends on the descriptors used, on the queried media collection and on the level of the user's similarity perception. In this work we endeavour to offer suitable distance measures for various situations. In future work, we will implement the distance measures identified as the best in the open MPEG-7-based visual information retrieval framework VizIR [11].

Acknowledgements

The author would like to thank Christian Breiteneder and Karyn Laudisi for their valuable comments and suggestions for improvement. The work presented in this paper is part of the VizIR project [11]. VizIR is funded by the Austrian Scientific Research Fund FWF under grant number P16111-N05.

References

1. Benchathlon network website (available from http://www.benchathlon.net/, last visited 2006-02-17)
2. Bober M (2001) MPEG-7 visual shape descriptors. Special issue on MPEG-7, IEEE Transactions on Circuits and Systems for Video Technology 11/6 : 716-719
3. Catell RB (1949) r_p and other coefficients of pattern similarity. Psychometrica 14 : 279-298
4. Chang SF, Sikora T, Puri A (2001) Overview of the MPEG-7 standard. Special issue on MPEG-7, IEEE Transactions on Circuits and Systems for Video Technology 11/6 : 688-695
5. Clark PS (1952) An extension of the coefficient of divergence for use with multiple characters. Copeia 2 : 61-64
6. Cohen J (1969) A profile similarity coefficient invariant over variable reflection. Psychological Bulletin 71 : 281-284
7. Czekanowski J (1913) Zarys metod statystycznych w zastosowaniu do antropologii. Prace Towarzystwa

Naukowego Warszawskiego 5

8. Del Bimbo A (1999) Visual information retrieval. Morgan Kaufmann, San Francisco CA
9. Eidenberger H (2003) Distance measures for MPEG-7-based retrieval. Proceedings ACM SIGMM International Workshop on Multimedia Information Retrieval, Berkeley CA: 130 - 137
10. Eidenberger H, Breiteneder C (2003) Visual similarity measurement with the Feature Contrast Model. SPIE volume 5021 (Storage and Retrieval for Media Databases Conference) : 64-76
11. Eidenberger H, Breiteneder C (2003) VizIR – a framework for visual information retrieval. Visual Languages and Computing 14 : 443-469
12. Eidenberger H (2004) Statistical analysis of visual MPEG-7 descriptors. ACM Multimedia Systems 10/2 : 84-97
13. Fuhr N (2001) Information Retrieval Methods for Multimedia Objects. In: Veltkamp RC, Burkhardt H, Kriegel HP (ed) State-of-the-Art in Content-Based Image and Video Retrieval. Kluwer, Boston, pp 191-212
14. Goodall DW (1967) The distribution of the matching coefficient. Biometrics 23 : 647-656
15. Gower JG (1967) Multivariate analysis and multidimensional geometry. The Statistician 17 : 13-25
16. Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bulletin Soc. Vaudoise Sciences Nat. 44 : 223-270
17. Jolion JM (2001) Feature similarity. In: Lew MS (ed) Principles of Visual Information Retrieval. Springer, Heidelberg, pp 121-144
18. Kulczynski S (1927) Bulletin International de l'Acadamie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles, Série B (Sciences Naturelles), Supplement II : 57-203
19. Lance GN, Williams WT (1967) Mixed data classificatory programs. Agglomerative Systems Australian Company Journal 9 : 373-380
20. Manjunath BS, Ohm JR, Vasudevan VV, Yamada A (2001) MPEG-7 color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technology 11/6 : 703-715
21. Manjunath BS, Salembier P, Sikora T (2002) Introduction to MPEG-7. Wiley, San Francisco CA
22. Meehl PE (1997) The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In: Harlow LL, Mulaik SA, Steiger JH (ed) What if there were no significance tests? Erlbaum, Mahwah NJ, pp 393-425
23. MPEG-7 eXperimentation Model website (available from http://www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html, last visited 2006-02-17)
24. MPEG-7 similarity measurement website (available from http://vizir.ims.tuwien.ac.at/SimEval, last visited

2006-02-17)

25. Ochiai A (1957) Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bulletin of Japanese Society for Science on Fish 22 : 526-530
26. Over P, Leung C, Ip H, Grubinger M (2004) Multimedia Retrieval Benchmarks. IEEE Multimedia 11/2 : 80-84
27. Pearson K (1926) On the coefficients of racial likeness. Biometrica 18 : 105-117
28. Rogers DJ, Tanimoto TT (1960) A computer program for classifying plants. Science 132 : 1115-1118
29. Russel PF, Rao TR (1940) On habitat and association of species of anopheline larvae in south-eastern Madras. Malaria Institute Journal 3: 153-178
30. Santini S, Jain R (1997) Similarity is a geometer. Multimedia Tools and Application 5/3 : 277-306
31. Santini S, Jain R (1999) Similarity measures. IEEE Transactions on Pattern Analysis and Machine Intelligence 21/9 : 871-883
32. Sint PP (1975) Similarity structures and similarity measures. Austrian Academy of Sciences Press, Vienna (in German)
33. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22/12 : 1349-1380
34. Sneath, P.H.A., and Sokal, R.R. Numerical Taxonomy. W. H. Freeman, San Francisco CA, 1973
35. Tversky A (1977) Features of similarity. Psychological Review 84/4 : 327-351
36. Web appendix of data tables (available from http://www.ims.tuwien.ac.at/~hme/papers/acmms04b-appendix-datatables-1.pdf, last visited 2006-02-17)
37. Webster H (1952) A note on profile similarity. Psychological Bulletin 49 : 538-539
38. Yule GU (1911) An introduction of the theory of statistics. Charles Griffin & Co., London UK

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

Table 1: Application of quantisation model on one-dimensional predicate vectors. The values of a, b, c, d depend on the predicate vectors X_i and X_j.

No.	*Measure*	*Comment*
P1	$a-\alpha b-\beta c$	Feature Contrast Model, Tversky 1977 [35]
P2	a	No. of co-occurrences
P3	$b+c$	Hamming distance
P4	$a+d$	Complement of Hamming distance [14]
P5	$\frac{a}{K}$	Russel 1940 [29]
P6	$\frac{a+d}{K}$	Simple match coefficient [14]
P7	$\frac{a}{a+b+c}$	Jaccard 1908 [16]
P8	$\frac{a}{b+c}$	Kulczynski 1927 [18]
P9	$\frac{a+d}{K+(b+c)}$	Rogers, Tamino 1960 [28]
P10	$\frac{2a}{2a+b+c}$	Czekanowski 1913 [7]
P11	$\frac{a}{a+2(b+c)}$	Sokal, Sneath 1963 [34]
P12	$\frac{(a+d)-(b+c)}{(a+d)+(b+c)}$	Hamann 1961 [32]
P13	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	Kulczynski 1927 [18]
P14	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c}\right)$	Sokal, Sneath 1963 [34]
P15	$\frac{a}{\sqrt{(a+b)(a+c)}}$	Ochiai 1957 [25]
P16	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Sokal, Sneath 1963 [34]
P17	$\frac{2(a+d)}{2(a+d)+(b+c)}$	Sokal, Sneath 1963 [34]
P18	$\frac{bc}{K^2}$	Pattern difference
P19	$\frac{ad-bc}{ad+bc}$	Yule 1911 [38]
P20	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Pearson 1926 [27]

Table 2: Predicate-based distance measures. See Subsection 3.3 for details.

No.	*Measure*	*Comment*
Q1	$\sum_k \left\| x_{ik} - x_{jk} \right\|$	City block distance (L^1)
Q2	$\sqrt{\sum_k \left(x_{ik} - x_{jk} \right)^2}$	Euclidean distance (L^2)
Q3	$\sum_k \frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}}$	Canberra Metric, Lance, Williams 1967 [19]
Q4	$\frac{1}{K} \sqrt{\sum_k \frac{\left(x_{ik} - x_{jk} \right)^2}{x_{ik} + x_{jk}}}$	Divergence coefficient, Clark 1952 [5]
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 [6]
Q7	$\frac{1}{K} \sum_k \frac{\left(x_{ik} - \mu \right)\left(x_{jk} - \mu \right)}{\sigma}$	Intra-class-coefficient, Webster 1952 [37]
Q8	$\frac{\sqrt{2K} - d_{ij}}{\sqrt{2K + d_{ij}}}$	Catell 1949 [3]
Q9	$\frac{\sum_k x_{ik} x_{jk}}{\sum_k x_{ik}{}^2 \sum_k x_{jk}{}^2}$	Angular distance, Gower 1967 [15]
Q10	$\sum_k^{K-1} \left(\left(x_{ik} - x_{ik+1} \right) - \left(x_{jk} - x_{jk+1} \right) \right)^2$	Meehl Index [22]

Table 3: Quantitative distance measures. See Subsection 3.3 for details.

Collection	*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
Coats-of-arms	9	12	Bavarian communal arms
	10	10	All Bavarian arms (difficult)
	11	18	Dark objects / light un-segmented shield
	12	14	Major charges on blue or red shield

Table 4: Ground truth information for Brodatz, Corel and coats-of-arms dataset.

	p		$p^{RETRIEVAL}$		$p^{BROWSING}$	
Descriptor	*Best*	*Ratio to MP7*	*Best*	*Ratio to MP7*	*Best*	*Ratio to MP7*
Color Layout	P19	67,1%	P19	29,6%	P6	67,3%
Color Structure	P19	61,2%	P19	18,5%	P18	66,1%
Dominant Color	P19	72,5%	P1	30,4%	P18	84,4%
Edge Histogram	P19	31,6%	P20	3,7%	P18	48,7%
Homogeneous Texture	P18	48,6%	P20	10,1%	P18	43,7%
Region-based Shape	P19	62,2%	P19	23,0%	P18	73,1%
Scalable Color	P6	65,5%	P19	48,6%	P4	59,6%

Table 5: Average relative performance indicator values of best distance measures related to MPEG-7 recommendations.

Ratios are defined as *performance indicator value for best measure / value for MPEG-7 distance measure* (percent).

Descriptor	*Measure 1*	*Measure 2*	*Ratio (p)*	*Ratio ($p^{RETRIEVAL}$)*	*Ratio ($p^{BROWSING}$)*
Edge Histogram	P1	P3	44,1%	12,4%	51,9%
Edge Histogram	P1	P8	46,3%	17,4%	51,4%
Edge Histogram	Q8	Q2	49,4%	39,8%	55,8%
Homogeneous Texture	P20	P16	89,7%	35,8%	98,5%
Scalable Color	P18	P20	84,5%	84,3%	84,6%

Table 6: Relative performance indicator values of selected distance measures for selected descriptors. Ratios are defined as in Table 5.

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

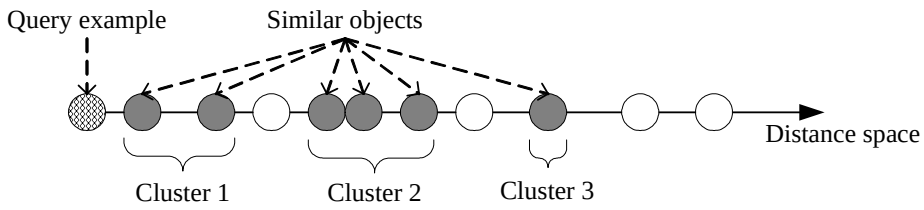


Figure 2: Example for clusters in distance space. The query example is positioned in the origin. Since distance measures map feature vectors to scalar values, distance space is always one-dimensional.

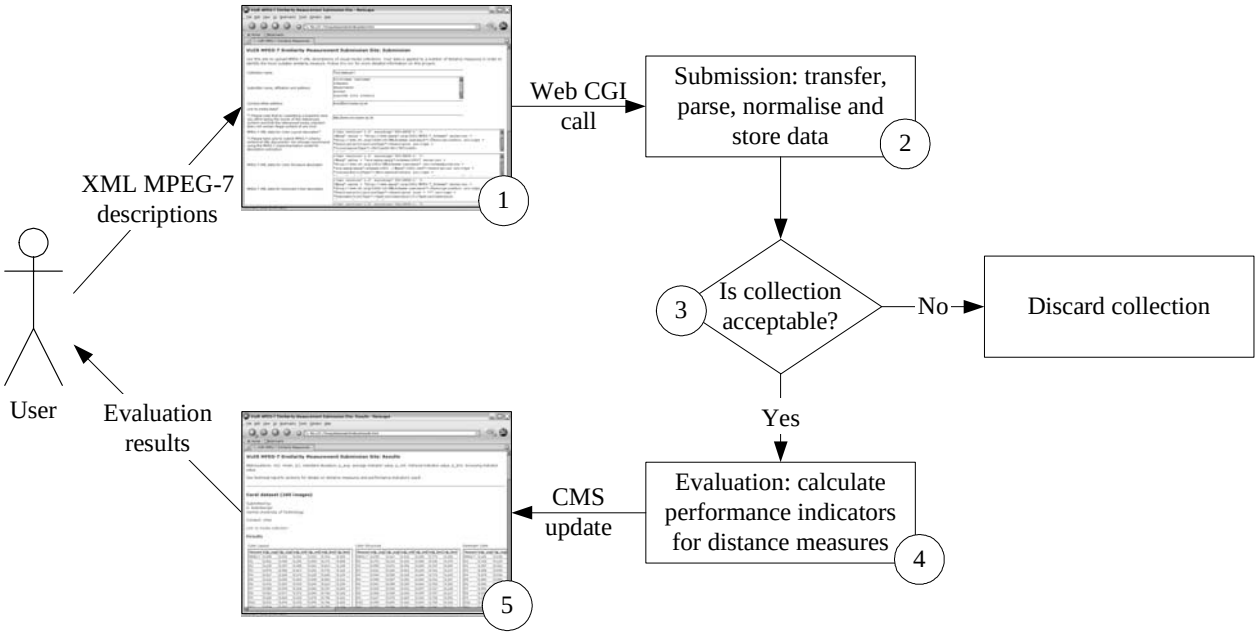
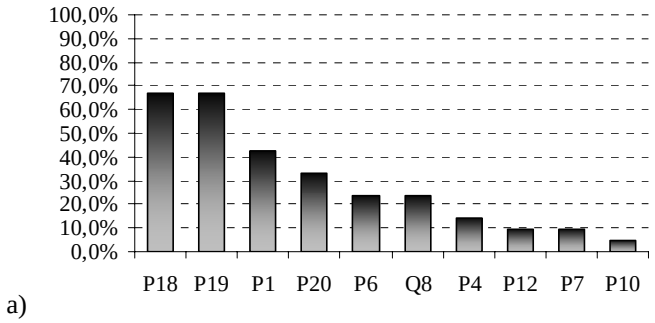
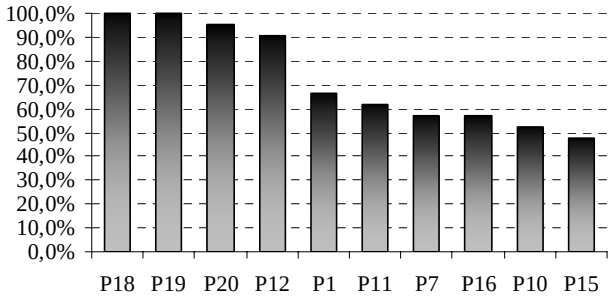


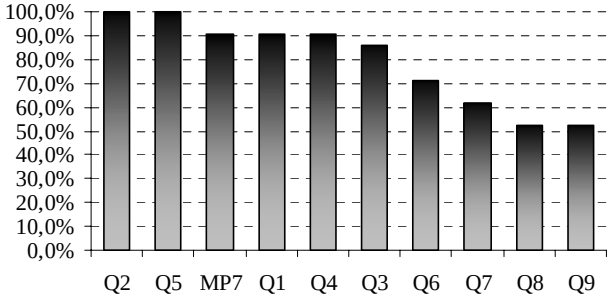
Figure 3: Workflow in public evaluation system. The user enters XML MPEG-7 descriptions in a web form. A server procedure computes the normalised data matrix and notifies the site manager. The site manager decides whether or not the submitted collection is acceptable. If yes, evaluation is performed and evaluation results are added to the results website.



a)

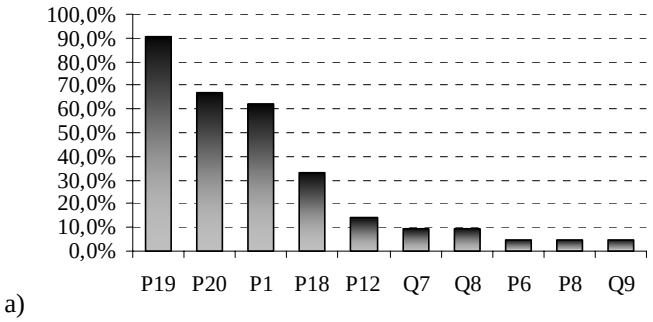


b)

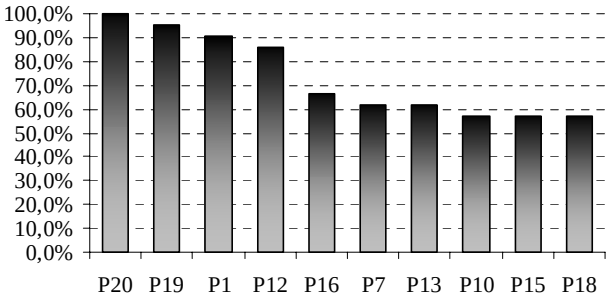


c)

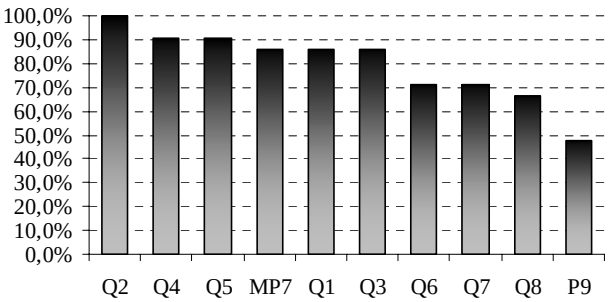
Figure 4: Best/worst ten distance measures ranked by performance indicator *p* (a: percentage of top 3 occurrences over all collections and descriptors, b: percentage of top 10 occurrences, c: percentage of occurrences among the worst 10 distance measures).



a)

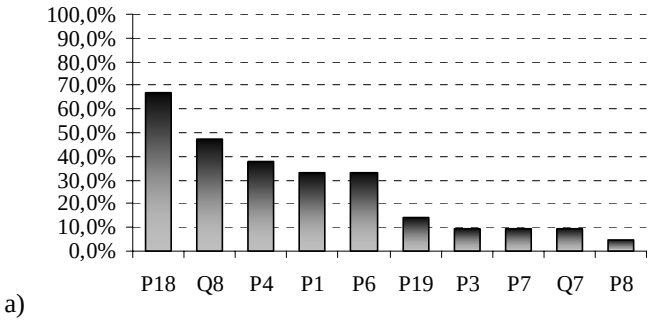


b)

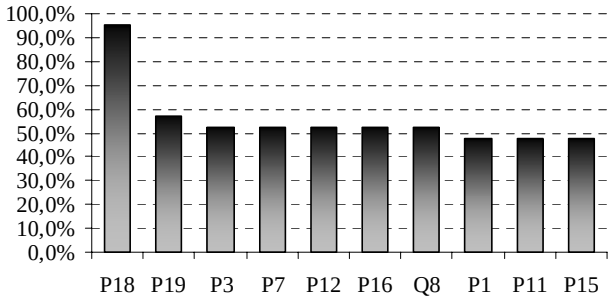


c)

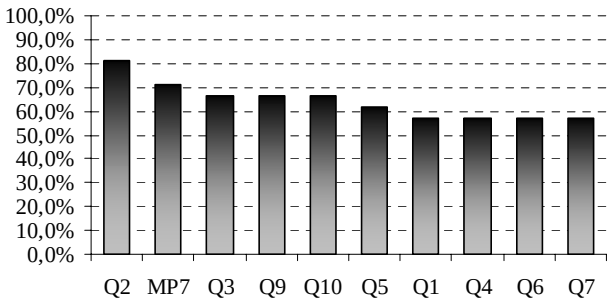
Figure 5: Best/worst ten distance measures ranked by performance indicator $p^{RETRIEVAL}$ (a: percentage of top 3 occurrences over all collections and descriptors, b: percentage of top 10 occurrences, c: percentage of occurrences among the worst 10 distance measures).



a)



b)



c)

Figure 6: Best/worst ten distance measures ranked by performance indicator $p^{BROWSING}$ (a: percentage of top 3 occurrences over all collections and descriptors, b: percentage of top 10 occurrences, c: percentage of occurrences among the worst 10 distance measures).

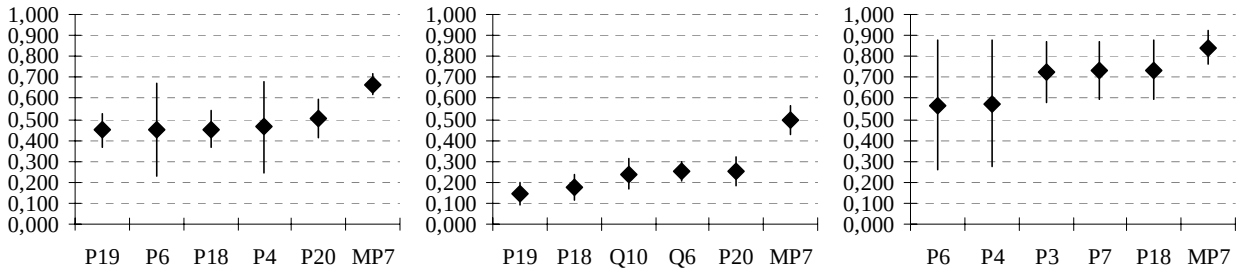


Figure 7: Performance of the best five distance measures and MPEG-7 recommendation on *Color Layout* descriptions (left: performance indicator p, middle: $p^{RETRIEVAL}$, right: $p^{BROWSING}$). The vertical axis shows the indicator values averaged over the three considered collections (diamond: mean, line: standard deviation).

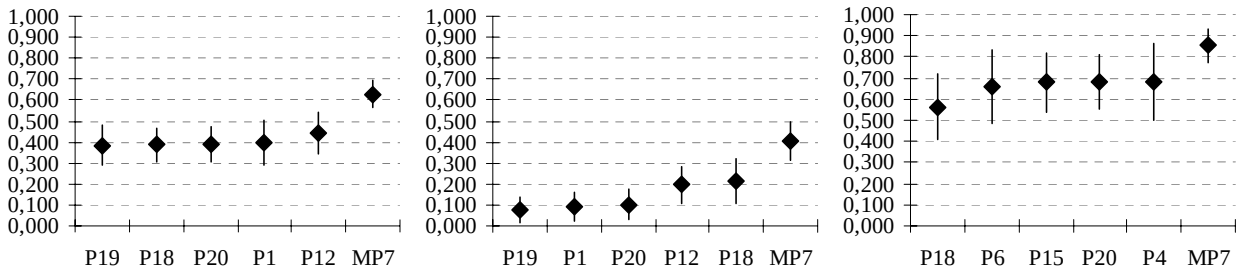


Figure 8: Performance of the best five distance measures and MPEG-7 recommendation on *Color Structure* descriptions (left: performance indicator p, middle: $p^{RETRIEVAL}$, right: $p^{BROWSING}$). The vertical axis shows the indicator values averaged over the three considered collections (diamond: mean, line: standard deviation).

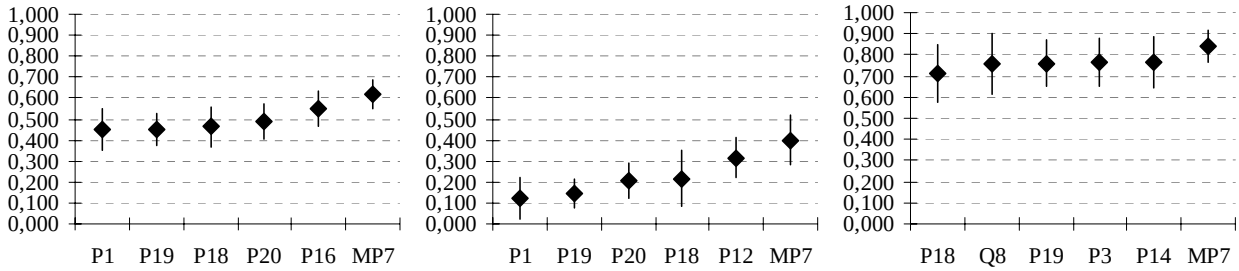


Figure 9: Performance of the best five distance measures and MPEG-7 recommendation on *Dominant Color* descriptions (left: performance indicator p, middle: $p^{RETRIEVAL}$, right: $p^{BROWSING}$). The vertical axis shows the indicator values averaged over the three considered collections (diamond: mean, line: standard deviation).

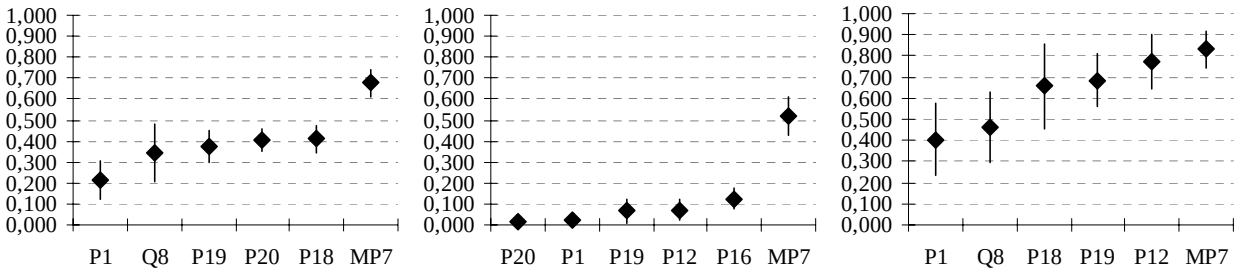


Figure 10: Performance of the best five distance measures and MPEG-7 recommendation on *Edge Histogram* descriptions (left: performance indicator p, middle: $p^{RETRIEVAL}$, right: $p^{BROWSING}$). The vertical axis shows the indicator values averaged over the three considered collections (diamond: mean, line: standard deviation).

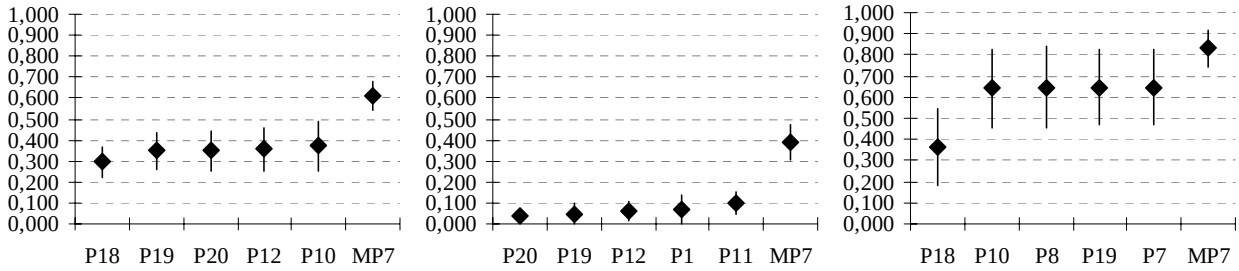


Figure 11: Performance of the best five distance measures and MPEG-7 recommendation on *Homogeneous Texture* descriptions (left: performance indicator p, middle: $p^{RETRIEVAL}$, right: $p^{BROWSING}$). The vertical axis shows the indicator values averaged over the three considered collections (diamond: mean, line: standard deviation).

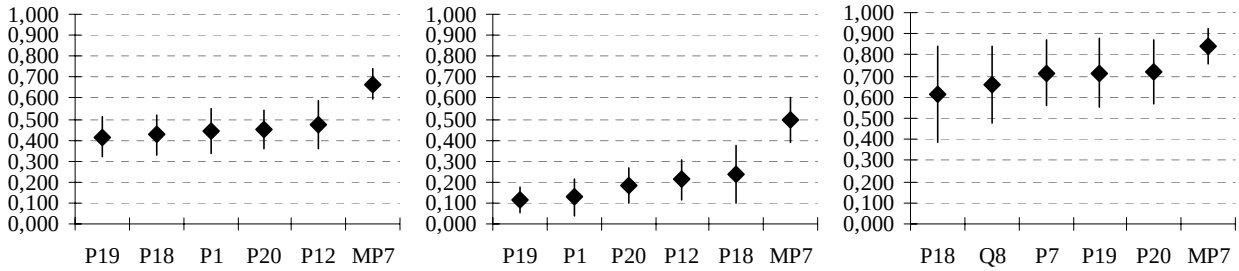


Figure 12: Performance of the best five distance measures and MPEG-7 recommendation on *Region-based Shape* descriptions (left: performance indicator p, middle: $p^{RETRIEVAL}$, right: $p^{BROWSING}$). The vertical axis shows the indicator values averaged over the three considered collections (diamond: mean, line: standard deviation).

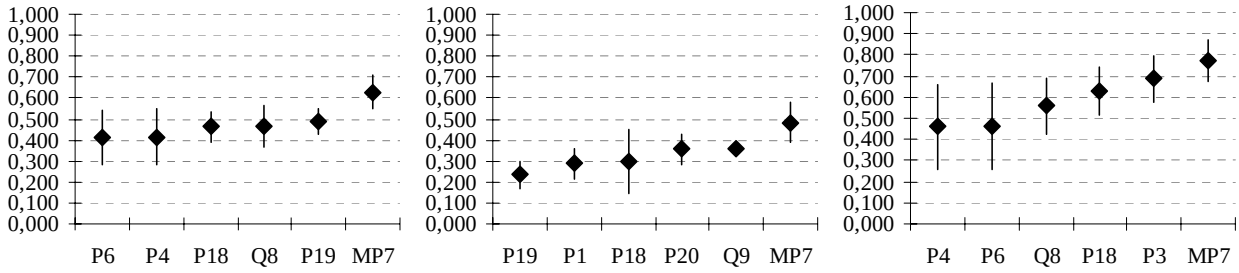


Figure 13: Performance of the best five distance measures and MPEG-7 recommendation on *Scalable Color* descriptions (left: performance indicator p, middle: $p^{RETRIEVAL}$, right: $p^{BROWSING}$). The vertical axis shows the indicator values averaged over the three considered collections (diamond: mean, line: standard deviation).