How good are the visual MPEG-7 features?

Horst Eidenberger[*]
Vienna University of Technology, Institute of Software Technology and Interactive Systems,
Favoritenstrasse 9-11, 1040 Vienna, Austria

ABSTRACT

The study presented in this paper analyses descriptions extracted with MPEG-7-descriptors from visual content from the statistical point of view. Good descriptors should generate descriptions with high variance, a well-balanced cluster structure and high discriminance to be able to distinguish different media content. Statistical analysis reveals the quality of the used description extraction algorithms. This was not considered in the MPEG-7-design process where optimising the recall was the major goal. For the analysis eight basic visual descriptors were applied on three media collections: the Brodatz dataset (monochrome textures), a selection of the Corel dataset (colour photos) and a set of coats-of-arms images (artificial colour images with few colour gradations). The results were analysed with four statistical methods: mean and variance of descriptor elements, distribution of elements, cluster analysis (hierarchical and topological) and factor analysis. The main results are: The best descriptors for combination are Color Layout, Dominant Color, Edge Histogram and Texture Browsing. The other are highly dependent on these. The colour histograms (Color Structure and Scalable Color) perform badly on monochrome input. Generally, all descriptors are highly redundant and the application of complexity reduction transformations could save up to 80% of storage and transmission capacity.

Keywords: MPEG-7, Statistical Analysis, Cluster Analysis, Factor Analysis, Self-Organizing Maps, Visual Information Retrieval, Content-based Image Retrieval, Content-based Video Retrieval

1. INTRODUCTION

The MPEG-7-standard[4] defines – among others – a set of basic descriptors for visual media content[7, 1]. In the design process these descriptors have been tested on large datasets with recall- and precision-like performance measures and ground truth information[7]. These tests give good indication on the performance of the descriptors in the retrieval process but say nothing on the *data quality* of the extracted descriptions (redundancy, distribution, etc.). The work described in this paper aims at assessing the quality of the extracted descriptions to give conclusions on the quality on the implemented description extraction algorithms. This is done by applying descriptors on pre-defined media collections and analysing the outcome with statistical methods. Surprisingly, to the author's knowledge this issue was not considered in the MPEG-7 design process.

Goal is to give guidelines, how descriptors should be used (e.g. for visual information retrieval[3, 8]) and how they could be improved. A good descriptor should fulfil several criteria. For example, it should be equally discriminant for any type of media (equal variance for different media parameters), the descriptor extraction process should be robust against different levels of quality and detail (all descriptor elements should contain reasonable data for any type of media), it should produce an equally distributed "net" of measurements over a well-defined media collection, etc. To test the MPEG-7-descriptors properties, statistical indicators as well as cluster analysis and factor analysis are used.

The paper is organised as follows. Section 2 gives background information on the MPEG-7-descriptors and the applied statistical procedures. Section 3 sketches the evaluation setup (including evaluation goals, used media collections and test environment). Section 4 contains the analysis results.

2. BACKGROUND

2.1 MPEG-7: visual descriptors

The visual part of the MPEG-7-standard defines several descriptors[7, 1]. 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.

[*] eidenberger@ims.tuwien.ac.at; phone 43 1 58801-18853; fax 43 1 58801-18898; www.ims.tuwien.ac.at

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 (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 localization 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 for 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 Methods for statistical data analysis

The major quality indicators for description extraction methods are the characteristics of the output descriptor *elements* (e.g. the bins of a color histogram). These are given as vectors over the size of the test dataset. The characteristics can be measured as variance *within* vectors, proximity *between* descriptor elements, distributions of quantised vector elements, etc. For the work described in this paper, five methods were used: (1) extraction of statistical indicators (mean and standard deviation) within vectors, (2) calculation of the distribution of quantised vector elements, (3) one-dimensional cluster analysis, (4) two-dimensional cluster analysis and (5) factor analysis. All of these methods are based on a single prerequisite: all descriptor elements have to measure on an interval-scale.

Mean and standard deviation give a general impression of a data vector. The mean characterises the average location of the underlying extraction method and the standard deviation gives a first clue on its discriminance. If the standard deviation is near zero, a description extraction method generates the same output for any type of given media content. Therefore, it may be characterised as non-discriminant. The second method from above quantises vector elements to a certain number of bins (e.g. ten) and aggregates the discrete distribution of the quantised values. The result is a density function that identifies the character of the extraction method (e.g. optimally, uniformly distributed) and helps finding holes within the measured range.

The cluster analysis methods derive groups of more than average similar description elements. Thus, redundancies in descriptors can be found. For one-dimensional analysis a hierarchical method was employed. The results of the analysis were presented by dendrograms. Next to clusters, the hierarchy of proximities can be used to assess the distribution of elements within the descriptor extraction space. The two-dimensional method enriches the result of the hierarchical cluster analysis by identifying clusters on a 2D-map. Such a map shows the relationships of clusters and may be used to identify *holes* in the measurement process. For the discussed analysis, Self-Organizing Maps[6] (SOMs) were used, because they produce a more natural clustering than, for example, k-means-clustering techniques. Finally, factor analysis was used as a method for elimination of redundancies in data vectors by identifying factors that cause the variance of the examined data. Additionally, factor analysis was used to identify common properties of descriptors by finding elements that load high on the same factor.

3. EVALUATION SETUP

Subsection 3.1 gives a complete and structured view on the research questions answered in this paper. Subsection 3.2 describes descriptors and media collections used. Subsection 3.3 gives a short sketch on the test environment and the parameters used in the analysis process.

3.1 Goals & questions

The major goal of this analysis is to give guidelines on how the visual MPEG-7-descriptors can be used on different media content and – if possible – to give suggestions how the description extraction process could be improved. In detail, the research questions addressed can be organised in three major groups: (1) descriptor design and usage, (2)

Group	Question	Statistical indicators
Descriptor design and usage	Which descriptors should or should not be used in combination?	Highly (un-)similar elements, high/low variance, highly/hardly discriminant, good/bad cluster structure
	How good is the net structure of the descriptors? Do they cover the whole range of variance?	Hierarchical clustering, topological clustering
	Where are holes in the extracted descriptions: for different types of media content and in general? In particular, do holes exist in the two major investigated descriptor-groups: colour and texture?	Non-uniform distribution of elements, empty clusters in maps
Dependency on media content	Which descriptors and descriptor elements perform well/badly on different types of content and in general?	High/low variance of elements, high/low discriminance
	Which content-related side effects on other (high-level) descriptors do exist?	
Redundancy identification	How redundant are the descriptors on different types of content and in general?	Number of factors, explained variance
	Which descriptors and descriptor elements address – for a specific type of content and in general – the same media characteristics?	High similarity, high self-similarity, common factors
	Which descriptor and descriptor elements could be compacted – for a specific type of content and in general – by transforming the data distribution without changing the characteristics of the extraction method?	Significantly deviated mean, highly non-uniformly distributed vectors
	How similar are – for a specific type of content and in general – the descriptors and descriptor elements of the two major investigated descriptor-groups: colour and texture?	Common factors

Table 1: Research questions and statistical indicators

dependency on media content and (3) redundancy identification.

The first group summarises guidelines for descriptor usage and improvement of description extraction. The second group, dependency on media content, analyses how sensitive MPEG-7-descriptors are to varying media content. This analysis allows conclusions on similarity measures for retrieval and on side-effects on other descriptors. The third group, redundancy identification, analyses the quality of MPEG-7-descriptions and investigates whether descriptions could be compressed. Table 1 summarises the research questions for the three groups and points out the statistical indicators used for clarification. The indicator values are the result of the analysis process.

3.2 Descriptors and media collections

Eight MPEG-7-descriptors were used for the statistical analysis. All colour descriptors: Color Layout (CLD), Color Structure (CSD), Dominant Color (DCD), Scalable Color (SCD), all texture descriptors: Edge Histogram (EHD), Homogeneous Texture (HTD), Texture Browsing (TBD) and one shape descriptor: Region-based Shape (RSD). The other basic shape descriptor, Contour-based Shape, was not used, because it produces structurally different descriptions that cannot be transformed to data vectors measuring on interval-scales. The motion descriptors were not used, because they integrate the temporal domain of visual media and would only be comparable if the basic colour, texture and shape

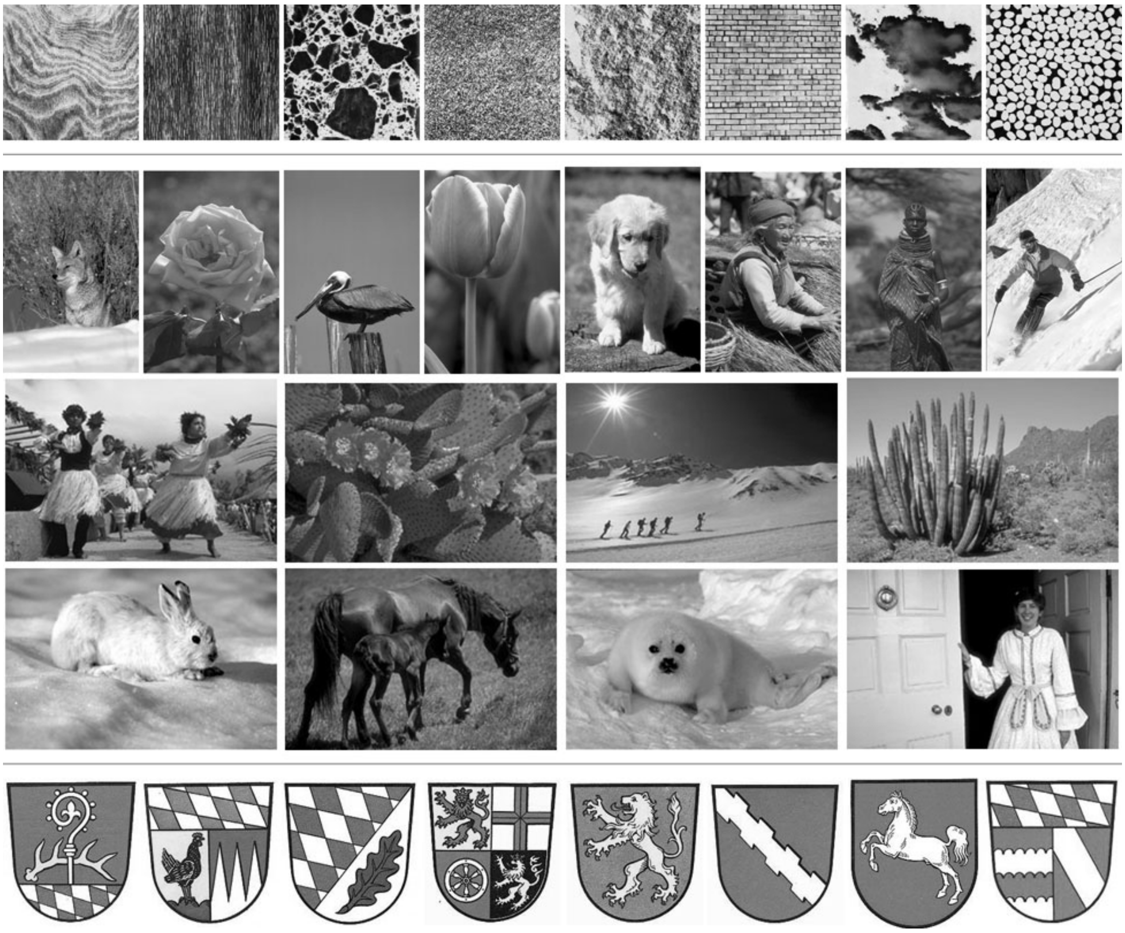

Figure 1: Test datasets. First row: Brodatz dataset, second to fourth row: Corel dataset, last row: coats-of-arms dataset.

descriptors would be aggregated over time. High-level descriptors were not 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.

The Texture Browsing descriptor had to be transformed to be useable in the evaluation. In the MPEG-7-standard it is defined as follows: (regularity, direction1, scale1, direction2, scale2) where regularity may be element of {not regular, slightly regular, regular, highly regular}, direction (in °) may be element of {no direction, 0, 30, 60, 90, 120, 150} and scale may be element of {no scale, fine, medium, coarse, very coarse}. Such a description is not suitable for the purpose of this paper. Therefore, the extracted descriptions were transformed to the following form: (regularity, scale: no direction, scale: 0, scale: 30, scale: 60, scale: 90, scale: 120, scale: 150) where regularity is element of {0 (not regular), 1 (slightly regular), 2 (regular), 3 (highly regular)} and the scale bins are element of {0 (no scale), 1 (fine), 2 (medium), 3 (coarse), 4 (very coarse)}. Such, all elements of Texture Browsing measure on an interval-scale.

Descriptor extraction was performed using the MPEG-7-reference implementation. In the extraction process each descriptor was applied on the whole content of each media object and the following extraction parameters were used. Colour in Color Structure was quantised to 32 bins (*ColorQuantSize=32*). For Dominant Color colour space was set to YCrCb (*ColorSpace=1*, *ColorSpacePresent=0*), 5-bit default-quantisation was used (*ColorQuantizationPresent=0*) and the default value for spatial coherency was used (*SpatialCoherency=0*). Homogeneous Texture was quantised to 32 components (*layer=0*). Scalable Color values were quantised to *sizeof(int)-3* bits and 64 bins were used (*NumberOfBitplanesDiscarded=3*, *NumberOfCoefficients=64*). Finally, Texture Browsing was used with five components (*layer=1*).

These descriptors were applied on three media collections with image content: the Brodatz dataset[3] (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[2] (426 images, 200x200 pixel). The Brodatz dataset is optimal for texture descriptors but a good test for colour and shape descriptors as well, because most colour descriptors are very sensitive for luminance and most shape descriptors use monochrome information for feature extraction. The Corel dataset (shipped with Corel Draw) is a widely-applied set of colour photos (showing humans, animals, flowers, landscapes, etc.) that has been used before to evaluate retrieval methods. For these evaluations a subset with images was taken from all collections. Especially, colour and texture descriptors should work well on this collection. The coats-of-arms dataset is somehow in-between these two collections: it consists of colour images with clear structures, few colour gradations and hardly any textures[2]. Therefore, colour and shape descriptors should work well on this dataset. Figure 1 shows examples from the three collections.

No collection from the MPEG-7-dataset was used in the evaluation, because the MPEG-7-descriptors were developed on the basis of these datasets. The evaluations should indicate, how well the descriptors perform on "unknown" material.

After the descriptor extraction, the resulting XML-descriptions were transformed into a data matrix with 798 lines (media objects) and 314 columns (descriptor elements). To be usable for statistical analysis the elements of this data matrix had to be normalised to a certain range. This was done for every element with a simple min-max-normalisation:

$$x'_{ik} = \frac{x_{ik} - min_k}{max_k - min_k} \tag{1}$$

where min_k is the minimum and max_k is the maximum for column k. The resulting value x'_{ik} is normalised to [0, 1]. This normalisation has the advantage that the distributions (variances) of both rows and columns of the data matrix are preserved.

3.3 Test environment and analysis parameters

The evaluation was performed in the following steps: (1) descriptor extraction, (2) XML-transformation and normalisation, (3) extraction of indicators, (4) quantisation and extraction of distributions, (5) hierarchical cluster analysis, (6) SOM calculation and (7) factor analysis. As pointed out above, the MPEG-7-reference implementation in version 5.6 was used (provided by TU Munich) to generate the descriptors. Image processing was done in Adobe Photoshop and normalisation and all other pre-processing steps (e.g. transformation of Texture Browsing descriptor) were computed with Perl. Hierarchical cluster analysis and factor analysis were calculated with SPSS and SOMs were calculated with SOM-PAK[6]. All other algorithms were implemented in Perl.

The following parameters were employed for the analysis: Mean and standard deviation were used as primary indicators for descriptor elements:

$$\mu_i = \frac{\sum_j^N x_{ij}}{N}, \sigma_i = \sqrt{\frac{\sum_j^N \left(x_{ij} - \mu_i\right)^2}{N}} \tag{2}$$

where x_{ij} are the extracted values of the i-th description element (column of the data matrix) and N is the number of investigated media objects. To identify the distribution of values of descriptor elements over N media samples the coefficients of the data matrix were quantised to ten bins. For the hierarchical cluster analysis a single-linkage algorithm with squared Euclidean distance measurement was used. The results were depicted as dendrograms on a relative scale from *0* (identical) to *25* (very un-similar).

SOMs were calculated with a hexagonal layout (every non-border cluster has six neighbours), 15 rows and 15 columns (225 clusters for 314 elements). For learning, a Gaussian neighbourhood kernel was used. Maps were initialised randomly. Learning was done in two iterations. In the first iteration 10000 learning steps were performed with learning rate $\alpha = 0.05$ and radius 10 (clusters). In the second iteration (fine tuning) 100000 learning steps were performed with learning rate $\alpha = 0.02$ and radius 3. For every dataset 15 separate SOMs were calculated and the best map was chosen by the minimum quantisation error, as defined by Kohonen et al.[6]. Because of the limited capacity of the SOM-PAK-implementation, only 200 (of 260) randomly chosen Corel images and 200 (of 426) coats-of-arms images could be used for training. See the SOM-PAK handbook for more information on the learning parameters.

For factor extraction a principal component analysis (analysis of the coefficients of the correlation matrix) was used. All Eigenvalues greater than one were selected as factors. To simplify interpretation, a Varimax-rotation was performed on the factor-loadings-matrix. Factor analysis can only be applied on elements with existing variance. Therefore, for the Brodatz dataset 225 elements could be used, for the Corel dataset 311 and for the coats-of-arms dataset 310.

4. RESULTS

Below, the results of the statistical analysis can be found. They are organised in groups and questions as introduced in Subsection 3.1. Because the questions on descriptor design and usage partially base on other questions the group order from Table 1 is reversed.

4.1 Redundancy identification

4.1.1 How redundant are the descriptors on different types of content and in general?

The redundancy of descriptors may best be identified by the results of a factor analysis. In general, a factor analysis has four major outputs: the number of extracted factors, the factors themselves, the amount of variance explained by each factor and in total and the loadings matrix that describes the influence of each factor on the analysed variables. Below, the first three outputs will be used to give general comments on redundancy in the investigated MPEG-7-descriptors.

For the Brodatz dataset 34 factors are able to explain 89% of the global variance. The first factor alone explains 15% of the variance. Since only those elements of the data matrix with an existing variance are used in the factor matrix, this result means that 34 factors explain 225 descriptor elements. This is a relationship of nearly 7:1. The MPEG-7-descriptors are highly redundant for monochrome media objects. This is not very surprising, because four of eight investigated descriptors are colour descriptors and some implemented colour descriptors work inferior on luminance information alone. Of course, this is a problem if MPEG-7-colour descriptors should be applied on media objects with monochrome content (e.g. for archival of old movies or drawings).

Applied on the Corel dataset, 69 factors explain 85% of the variance of all descriptor elements. The first factor explains 12%. With 311 input columns from the data matrix this means that – applied on coloured content with rich details – the MPEG-7-descriptors are redundant with a ratio of about 9:2. This is surprising because the Corel photos include brilliant colours, significant edges and textures and – many of them – tailor-made object arrangements for the Region Shape descriptor. The reasons for this high redundancy will be identified in the answers on the questions below.

The factor analysis on the coats-of-arms dataset extracts 71 factors. These factors are able to explain 80% of the global variance. The best factor is able to describe just 6.7% variance. 310 elements were analysed. Therefore the ratio of redundancy is again 9:2. It is surprising that the MPEG-7-descriptors perform slightly worse on the coats-of-arms dataset than on the Corel dataset: the Corel photos contain much more details and, generally, descriptors should be less redundant on material with more content. The reason for the opposite results may be that the characteristics of the coats-of-arms dataset are positioned between the Brodatz and the Corel collection. Consequently, the results of the factor analysis suggest that the coats-of-arms may be a better context-free test than the two other collections.

In sum, the MPEG-7-descriptors generate results of high redundancy (4:1 to 7:1). If the MPEG-7-descriptors are used in the situation where storage is limited, it may be a good idea to make use of transformations for data compression (e.g. KLT[3]). Additionally, this analysis gives a first hint that MPEG-7-colour descriptors have problems with monochrome content.

4.1.2 Which descriptors and descriptor elements address – for a specific type of content and in general – the same media characteristics?

To answer this question it will be investigated, which descriptors generate the same output for the same input. The results can be used to identify redundancies in the extraction process but, as well to give guidelines for descriptor usage in descriptor schemes (see 4.3.1). Useful statistical information is provided by clustering algorithms (proximities in the analysis results, cluster distributions) and factor analysis (common factor loadings of elements).

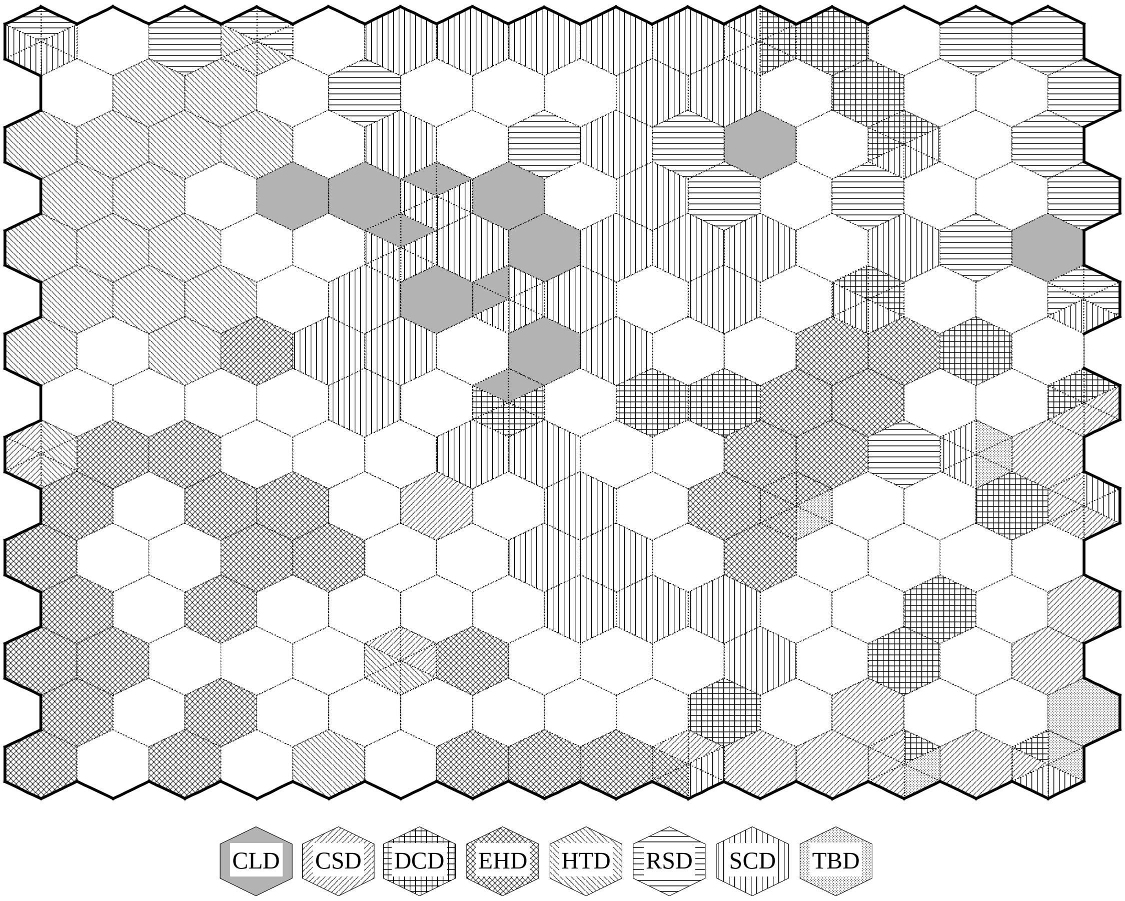

Figure 2: Self-Organizing Map for Corel dataset (CLD: Color Layout, CSD: Color Structure, DCD: Dominant Color, EHD: Edge Histogram, HTD: Homogeneous Texture, RSD: Region Shape, SCD: Scalable Color, TBD: Texture Browsing).

The best starting point for the analysis is the hierarchical cluster structure. It shows intuitively proximities between elements (even though dendrograms become quite large for 314 elements). A first striking result is the high self-similarity of the elements of Homogeneous Texture for any type of media. For the Brodatz dataset (rich textures) and the coats-of-arms dataset (poor textures) all elements of this descriptor form a single cluster with a maximum distance of 1 (scaled on 0-25 where 25 is the distance value for a vector of 0-values to a vector of 1-values). For the Corel dataset the descriptor forms two clusters where the larger one has the same characteristics as for the other two media sets. Only some energy values (no energy deviation) fall apart and form a cluster with distance 5 to the main cluster. In conclusion, the elements of the Homogeneous Texture descriptor are – independent of the media – highly self-similar and redundant.

All other elements form clusters with a distance of at least 3. Interesting is the Edge Histogram that forms 5 to 10 clusters with 10 to 15 elements of distance 3 to 5 for any type of content. The elements of these clusters are self-similar but the distance between the clusters is relatively large. The behaviour of this descriptor (5-bin-edge histograms for predefined regular regions) could be described as taking a certain quantity of samples from the media and describing each sample with a set of similar elements. Additionally, some elements of Region Shape and Scalable Colour form smaller clusters but most elements are – from the one-dimensional point of view – not very redundant.

Looking at the SOM-clustering gives a "topological" view on the data. Working on a map, the clustering algorithm has two degrees of freedom to arrange elements. Therefore, *relationships* between elements can more easily be visualised and some elements may be grouped closer to each other than they would in a one-dimensional cluster analysis.

Analysing the SOMs for the three media collections (Figure 2 shows the map for the Corel dataset) supports the first impression of the hierarchical analysis. Homogeneous Texture lays a fine-meshed net over the investigated media property. Independent of the media content it consists of 15 to 20 clusters with 3 to 5 elements each. These clusters form a homogeneous super-cluster within the map: each cluster is connected to at least one other cluster and the border of the super-cluster is nearly circular. Therefore, Homogeneous Texture is really homogeneous but the measured property could be expressed with much less description elements.

The Edge Histogram descriptor forms 2D-clusters with slightly more elements than the Homogeneous Texture descriptor. These clusters are spread over large regions and only loosely connected. As described above, the net of the Edge histogram is wide-meshed but the descriptor covers a larger area of the media content. An interesting phenomenon can be observed for the Brodatz dataset. On such content, the descriptor should be highly discriminant but in the map most Edge Histogram clusters are border clusters. This suggests that descriptions generated by the Edge Histogram are extreme (in terms of variance) and not as discriminant as they should be.

For any content all other descriptors form rather small 2D-clusters. These clusters are wide-spread over the whole maps. It may be concluded that most descriptors implement different ideas. Especially, the four colour descriptors are astonishingly non-overlapping.

A more detailed view than the cluster analysis can be derived from the factor loadings matrices of the factor analysis. The coefficients of these matrices express to which extent a factor influences the variables (elements) from which it is derived. Elements that have a high load on the same factor should have a similar variance. In the following, in the first step the most important factors for each type of media will be investigated and in the second step other interesting factor correlations will be pointed out.

The first factor for the Brodatz dataset (explains 15%) loads heavy on the DC-value for the Y-colour-component of Color Layout, the second colour bin of Dominant Color (the first dominant colour is probably white), half of the energy values and deviations of Homogeneous Texture, almost all bins of Region Shape and bin 7, 9, 13 of Scalable Color. The latter bins are probably responsible for greyscale pixels. These elements are highly similar and, for example, the Y-DC-coefficient could be used as a good indicator for them. Factors 2, 4-6 and 8 (9%, 8%, 7%, 5%, 4%) measure on the five edge types of the Edge Histogram. The values for each edge type are highly similar and using a global edge histogram (as defined in the standard) could be a good idea on content comparable to the Brodatz dataset. The third factor (8%) loads on the half of the elements of Homogeneous Texture that is not explained by the first factor. These bins (mainly 4-10, 19-24) are highly redundant.

The first factor for the Corel dataset (explains 12%) loads high on non-directional edges and almost all elements of Homogeneous Texture. This supports the impression that Homogeneous Texture is highly redundant. Similarity of non-directional edges is easily explainable from the applied extraction algorithm. The second factor (7%) loads high on the Y-DC-coefficient of Color Layout and most elements of Region Shape (as for the Brodatz dataset). Surprisingly, it seems that the first colour bin and all Region Shape elements are highly correlated independent of the complexity of the media content. Other factors do not show significant correlations of elements. For example, for complex media content it seems that edge bins of the Edge Histogram are very different from each other. Therefore, using a global histogram on complex content may not be a good idea.

For the coats-of-arms database the first factor (7%) loads high on non-directional edges and half of the Homogeneous Texture bins. Even though these media objects hardly contain texture information, non-directional edges and energy-bins are highly correlated. This allows the conclusion that these elements are highly redundant and non-directional edges may be used as a substitute for Homogeneous Texture. Looking at the cluster analysis results reveals that, indeed, these elements are clustered close to each other. Factor 2 and 3 (4%) load high on various colour bins of the colour descriptors. This is not surprising as coats-of-arms images mainly contain large coloured regions. A significant correlation of bins of the same edge type could not be identified.

Another interesting result of the factor analysis is that – for any type of content – the Dominant Color descriptor has the tendency to identify colours with the same colour-component values. According to the used parameters dominant colours are described in the YCrCb-colour space. Somehow, the Y-, Cr-, and Cb-component of a dominant colour are most times highly similar. Maybe there is a certain characteristic in the extraction algorithm that causes this phenomena. The conclusion has to be that the Dominant Color descriptor is highly sensitive to brightness and, indeed, it is the best colour descriptor for greyscale objects (see 4.2). Another interesting result is that all bins of Color Layout are for any type of media highly un-similar. In every map and every loadings matrix a separate cluster respective factor can be found for any element of Color Layout.

To conclude this question, the elements of Homogeneous Texture are highly redundant, Edge Histogram consists of clusters of redundant elements, non-directional edges explain Homogeneous Texture and Dominant Color is very sensitive to brightness.

4.1.3 Which descriptor and descriptor elements could be compacted – for a specific type of content and in general – by transforming the data distribution without changing the characteristics of the extraction method?

Ideally, each description element should have a uniform distribution over all elements of a sufficiently large media collection. Such description elements would be optimally discriminant and data space would be better utilised. In reality, most descriptors have holes: ranges of result values that are not covered. In the following it will be investigated if the MPEG-7-descriptors have such holes by analysing the statistical indicators and the distributions of elements.

The Color Layout descriptor tends to not use the lowest 10% of all possible values. For the Brodatz dataset, all values are higher than 0.3 (values are normalised to [0, 1]!), the average mean is about 0.7 and the standard deviation is lower than 0.1. For Corel and coats-of-arms dataset the situation is less bad but still not optimal. Values below 0.1 never occur. All other values are sufficiently utilised. That means, Color Layout values could be transformed and quantised to a smaller data type. This could save (depending on the content) at least 10% of storage space.

An interesting phenomenon can be observed for the Edge Histogram. Independent of the type of media this descriptor does not measure on the ranges [0.32, 0.42[, [0.72, 0.81[and [0.9, 1[. Because this descriptor simply counts edges of certain orientations in pre-defined spatial regions the reason for this behaviour must lie in the extraction process itself (maybe in the edge operators?). In any case, this phenomenon allows extended compacting of the descriptor. About 30% of the allowed data range are not used. With a suitable transformation and quantisation the needed storage and transmission space for this descriptor could be reduced to 70%.

Most energy values and deviations of Homogeneous Texture measure exclusively on the range [0.5, 1]. The mean of nearly all bins is about 0.7 and standard deviation is about 0.1. Except for bins 12 to 15 for colour photos the first half of values is not utilised. Therefore, with an appropriate transformation and quantisation the amount of needed storage for Homogeneous Texture could be reduced to 50%.

The same phenomenon as for Edge Histogram can also be observed for Scalable Color. For any type of media (except the first 16 colour bins) the ranges [0.1, 0,2[, [0.4, 0.5[and [0.7, 1[are not used at all. This may have to do with the application of the Haar transformation. In any case, about 45% of storage capacity could be saved by compacting this descriptor.

The other descriptors do not show holes or just for the Brodatz dataset that represents an extreme case of dataset. In conclusion, if any element of a complete MPEG-7-description is stored as an 8 byte Double, the amount of needed storage could be reduced from 2512 byte to 1832 byte (73%) without losing precision in the data.

4.1.4 How similar are – for a specific type of content and in general – the descriptors and descriptor elements of the two major investigated descriptor-groups: colour and texture?

In the following the main similarities that could be identified by cluster and factor analysis are described. As mentioned above, for all types of media the DC-coefficient of the luminance of Color Layout determines most elements of Region Shape. This element seems to be a good identifier – even for complex scenes – for global shape information (as Region Shape should be).

The colour descriptors are similar in the sense that they perform badly if just one colour component is present. An exception is Dominant Color. Consequently, this descriptor is absolutely independent of all other colour features and shows no similarities to texture descriptors either. Significant similarities among the colour histogram descriptors could not be identified. Maybe because they work on different colour spaces they generate absolutely independent results.

Within the texture group, the non-directional edges describe the content of Homogeneous Texture astonishingly precisely. The third texture descriptor, Texture Browsing, is independent of all other texture descriptors. Texture

Browsing does not show similarities to any other descriptor.

4.2 Dependency on media content

4.2.1 Which descriptors and descriptor elements perform well/badly on different types of content and in general?

As defined above, a good descriptor should have a high variance and produce description values that are uniformly distributed over the whole output range. These properties can be measured with statistical indicators (mean and standard deviation) and the distribution analysis. Additionally, looking at the cluster structure of the hierarchical cluster analysis may help to judge the discriminance of descriptors. Below, the results for all investigated descriptors and the three tested media collections can be found.

Color Layout performs badly on monochrome data. Only six of twelve bins have a standard deviation greater than null: the DC- and the AC-coefficients of the luminance channel. Even for these, the standard deviation is only about 0.1 and the mean is above 0.7. For the Corel database the descriptor works well: mean is about 0.5 and standard deviation is about 0.2. For content with less colour gradations the performance drops again: the standard deviations goes down to 0.1 and the mean up to 0.7. For any type of media, all elements with existing variance are Gaussian-distributed and of suboptimal discriminance.

Like Color Layout, Color Structure performs inferior on monochrome data: 24 of 32 colour bins have no variance, the other bins are apparently responsible for brightness (bin 37 to 44). Even these bins have a very low variance. Whenever colour is present – independent of the number of gradations – Color Structure performs good. The mean is mostly below 0.5 but the standard deviation is good: in average 0.25. Therefore the element values are distributed over the entire range of possible values. The distribution of most elements is irregular. Only a few elements are Gaussian-distributed.

The Dominant Color identifier performs equally well on any type of media. The mean of generated values is usually a bit lower than 0.5 and the standard deviation is about 0.3. Therefore the utilisation of the domain is excellent. If only few colours or brightness values are available the standard deviation for the last dominant colours (five bins in total) drops down to zero. The distribution of values is very similar for any type of media. Most percentage-bins are Gaussian-distributed while the colour bins are mostly irregularly distributed. This may have to do with the phenomena described above that the colour component values seem to be coupled. Only a few elements of the descriptor are nearly uniformly distributed.

The last colour descriptor, Scalable Color, performs exactly like Color Layout and Color Structure. For monochrome content, it is unable to derive meaningful descriptions. Only eight bins have an existing variance (bin 1-3, 5, 9, 13, 33 and 49) but these elements have a mean of about 0.4 and an average standard deviation of 0.3. For the Corel dataset the results are excellent: mean 0.5 and standard deviation 0.3. For less colour gradations than in photos many standard deviation values drop down below 0.1. This means, for non-photo-content the descriptor is hardly discriminant. Additionally, nearly all elements have an irregular distribution over the ten investigated bins.

The results for colour descriptors can be summarised as follows. All colour descriptors work excellent on photos but three of four perform badly on artificial media objects with few colour gradations and very badly on monochrome content. An exception is Dominant Color. This descriptor works well on each type of content. A solution for using colour descriptors on media objects with a single colour component could be storing, transporting and utilising only the bins that are sensitive for brightness (e.g. using special distance measures for retrieval that take only these elements into account).

Edge Histogram performs excellent on any type of media. Even on coats-of-arms images with hardly any textures present (but, of course, very clear edges) the mean of values is near 0.5 and most elements have a standard deviation higher than 0.25. For different content, the statistical indicators are even better. The problem observed in Subsection 4.1.2 could not be verified: the Edge Histogram descriptor does not produce extreme results on texture images. Because of the phenomenon described in Subsection 4.1.3 (holes in the distribution) all elements are irregularly distributed. Most have two significant peaks between [0, 0.1] and [0.8, 0.9].

The Homogeneous Texture descriptor works acceptably on the Brodatz dataset (even though the standard deviation of most elements is below 0.1) but relatively poor on colour images, especially if they have few colour gradations and textures in them. In this case the standard deviation of most elements drops below 0.05. This means, Homogeneous Texture is not discriminant for colour media and still relatively poor for texture regions. The mean of most elements is about 0.75. Nearly all descriptor elements are – for any content – Gaussian-distributed and because the standard deviation is very low, only a few bins are utilised.

Texture Browsing, in the form described in Subsection 3.2, performs good on the Brodatz dataset and the Corel dataset (standard deviation about 0.2) but poorly on the coats-of-arms dataset (standard deviation lower than 0.05). The means of all elements are significantly below 0.5 and all elements are irregularly distributed. Because of the poor variances and the few elements of this descriptor (eight bins) it is difficult to use in most applications, because results may be ambiguous (e.g. retrieval).

The conclusion over all texture descriptors must be that Edge Histogram is – from the statistical point of view – by far the best descriptor. Homogeneous Texture is highly sensitive on the analysed media content and the variance of results is small. Texture Browsing produces partially ambiguous results that are, indeed, suitable for browsing but not for other MPEG-7-applications.

The last descriptor is Region-based Shape. This descriptor measures excellently on any type of media. The mean of most elements is lower than 0.5 but the standard deviation is in average 0.2 to 0.25. Most elements are uniformly distributed. The other are Gaussian distributed but over a wide range. Therefore, Region-based Shape is a good descriptor in any situation that can be applied to any type of media content.

Finally, looking at the hierarchical cluster structure reveals that – for any type of media – hardly any clusters exist on average distance (5-15). Most clusters are on distances lower than 5. The size of these clusters varies widely. This allows the conclusion that the elements of most descriptors are far from being uniformly distributed (as they should be). If they were, a lot of clusters with similar sizes should exist on average level and the dendrograms should look like well-balanced trees.

4.2.2 Which content-related side effects on other (high-level) descriptors do exist?

These insights cause one major side-effect on other descriptors. The Group-of-Frames/Group-of-Pictures descriptor (GoF/GoP) is based on Scalable Color. It is intended to generate descriptions for video clips and animations. The description is calculated by taking the mean over all colour histograms of media objects in a group. Obviously, this descriptor does not work if the given media does not consist of nicely shot pictures with lots of colour gradations. For example, it cannot be used for animations, old monochrome movie clips, cartoons, etc. The same would be true if Color Layout or Color Structure was used.

From the statistical results it has to be concluded that GoF/GoP should be based on Dominant Color instead. For monochrome content the implementation would be straight-forward, because colours based on one component can easily be averaged. For colours with three components the implementation would be a bit more complicated. Still, the problem is solvable if additional knowledge on colour models is integrated (to find corresponding colours) in the averaging process.

4.3 Descriptor design and usage

4.3.1 Which descriptors should or should not be used in combination?

The answer to this question is partially based on the examinations from above (un-similar elements, elements with high variance, etc.). Additionally, factor analysis can give valuable information. If a factor loads high on two elements but with opposite signs (factor loadings are element of [-1, 1]) then these elements are highly un-similar and suitable for being used in combination. The same is true for elements that are caused by factors that do not load on any other element.

The evaluations so far have revealed that Dominant Color, Edge Histogram and Texture Browsing are the most independent descriptors. Color Layout is independent of other descriptors as well and to large proportions the Y-DC-bin determines the Region Shape descriptor. The three major colour descriptors, Color Layout, Color Structure and Scalable Colour, are independent of each other and other descriptors as well, if enough colour information is present in the content. If not, they perform equally badly and their results become very similar. This means, on non-photographic content it does not make sense to combine these descriptors. Homogeneous Texture is to a large extent determined by the Edge Histogram descriptor. From the statistical point of view it does not make sense to combine these descriptors.

This is supported by additional analysis of the factor loadings matrices. Color Layout, Dominant Color, Edge Histogram and Texture Browsing are mainly explained by unique factors (even though Dominant Color and Edge Histogram contain a certain amount of self-similarity, as described above). If relationships exist, they are highly negative (e.g. the AC-coefficients of Color Layout and all elements of Texture Browsing are pair-wise highly negatively loaded). Most elements of the other descriptors depend on factors of these descriptors.

In conclusion, the ideal – content-independent – descriptor scheme for visual content seems to be Color Layout (because of the luminance coefficients), Dominant Color, Edge Histogram and Texture Browsing. Still, the other descriptors may be meaningful in specific situations and application scenarios.

4.3.2 How good is the net structure of the descriptors? Do they cover the whole range of variance?

This question can be answered best by the results of clustering. A good descriptor should span over the whole range of possible values (e.g. for a certain media class over the whole SOM) and consist of small clusters. In this case, the outcome would be uniformly distributed and similar and un-similar media objects would be represented as they should be. Unfortunately, in reality such a net structure does hardly exist.

As described above, Homogeneous Texture spans a fine-meshed net but only over a small area of the available variance (and therefore, for the price of high redundancy). Edge Histogram spans a large wide-meshed net with large clusters. This bears the risk that media objects that are only slightly different may be – in one case – described equally and in another case completely different. On the other hand the descriptor is generally less redundant. Scalable Color generates a mesh that is very similar to Edge Histogram: larger clusters and widely meshed. In the author's opinion using this descriptor with more bins (e.g. 256) would not improve the net structure because of the large clusters. The outcome of Region Shape is mostly widely meshed with large clusters. Finally, Color Structure has smaller clusters but very large holes in the net structure. For this descriptor, using more colour bins could improve the net structure and reduce the size of the holes (or increase the size of clusters). The other descriptors have too few elements to be able to speak of a net structure.

4.3.3 Where are holes in the extracted descriptions: for different types of media content and in general? In particular, do holes exist in the two major investigated descriptor-groups: colour and texture?

This question aims at finding holes in the generated cluster structure that are not covered by any descriptor and therefore, cannot be filled with descriptor schemes either. Most easily, such holes can be found with the cluster analysis results, especially the 2D-clusterings.

Looking at the SOM for the Brodatz dataset shows that large holes exist between the Homogeneous Texture descriptor and the Edge Histogram descriptor. One reason for this is that the colour histogram features that fill these holes for other content do not work on monochrome content. Therefore, two solutions are possible to close these holes: changing the algorithms of the colour histograms to make them sensitive for greyscale media objects or modifying the neighbouring descriptors, especially Edge Histogram. Edge Histogram should produce a more homogeneous cluster structure with smaller clusters. This could be achieved by making the extraction mechanism more sensitive for small differences in the content (e.g. by using less coarse edge detectors).

For the Corel dataset many small holes exist. The large holes are filled by elements of Scalable Color. This means, if the descriptor is used in combination with the best other descriptors for this type of data, then the resulting descriptions have a fine-meshed structure and even small differences in the content are recognisable by the descriptions. For the coats-of-arms dataset the same problem occurs as for the Brodatz dataset. Large holes can be found between Edge Histogram and Homogeneous Texture, because Scalable Color is missing. The cure should be the same as for the Brodatz dataset. Scalable Color should be more sensitive if only few colour gradations are present.

In summary, nearly no holes exist between clusters of colour descriptors. If they work, these descriptors are very independent from each other and generate a fine-meshed structure with clusters of – at most – average size. For the texture descriptors large holes exist between Edge Histogram and Homogeneous Texture if Scalable Color does not work. Texture Browsing is – because of its poor variance – unable to close these holes.

5. CONCLUSIONS

The study presented in this paper analyses descriptions extracted with MPEG-7-descriptors from visual content from the statistical point of view. Good descriptors should generate descriptions with high variance, a well-balanced cluster structure and high discriminance to be able to distinguish different media content. Statistical analysis reveals the quality of the used description extraction algorithms. This was not considered in the MPEG-7-design process where optimising the recall was the major goal. For the analysis eight basic visual descriptors were applied on three media collections: the Brodatz dataset (monochrome textures), a selection of the Corel dataset (colour photos) and a set of coats-of-arms images (artificial colour images with few gradations). The results were analysed with five statistical methods: mean and variance of descriptor elements, distribution of elements, cluster analysis (hierarchical and topological) and factor analysis.

The main results are: The best descriptors for combination are Color Layout, Dominant Color, Edge Histogram and Texture Browsing. The others are highly dependent on these. The colour histograms (Color Structure and Scalable Color) perform badly on monochrome input. Therefore, Dominant Color should be used for GoF/GoP colour instead of Scalable Color. Generally, all descriptors are highly redundant and applying complexity reduction transformations could save up to 80% of storage and transmission capacity. Finally, some aspects of images are captured by none of the descriptors and existing descriptors should be either refined or new visual descriptors be added to the standard.

The study was prepared for the visual information retrieval project VizIR[5]. VizIR is based on the visual MPEG-7-descriptors and, as a consequence of the results, the future implementation focus will lay on Color Layout, Dominant Color, Edge Histogram and Texture Browsing.

ACKNOWLEDGEMENTS

The author would like to thank Christian Breiteneder for his valuable comments and suggestions for improvement. Additionally, this work and the VizIR project are supported by the Austrian Scientific Research Fund (FWF) under grant no. P16111-N05.

REFERENCES

1. M. Bober, "MPEG-7 Visual shape descriptors", *Special Issue on MPEG-7, IEEE Transactions on Circuits and Systems for Video Technology*, **11/6**, 716-719, 2001.

2. C. Breiteneder, H. Eidenberger, "Content-based Image Retrieval of Coats of Arms", *Proceedings IEEE International Workshop on Multimedia Signal Processing*, 91-96, IEEE, Helsingör, 1999.

3. A. Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann Publishers, San Francisco , 1999.

4. S.F. Chang, T. Sikora, A. Puri, "Overview of the MPEG-7 standard", *Special Issue on MPEG-7, IEEE Transactions on Circuits and Systems for Video Technology*, **11/6**, 688-695, 2001.

5. H. Eidenberger, C. Breiteneder, "A Framework for Visual Information Retrieval", *Proceedings Visual Information Systems Conference*, 105-116, Springer Verlag, HSinChu 2002.

6. T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, *SOM-PAK: The Self-organizing Map Program Package*, Technical Report, Helsinki University of Technology, 1995.

7. B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, A. Yamada, "Color and texture descriptors", *Special Issue on MPEG-7, IEEE Transactions on Circuits and Systems for Video Technology*, **11/6**, 703-715, 2001.

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