
Semantics in Content-based Multimedia Retrieval

Horst Eidenberger¹, Maia Zaharieva²

Vienna University of Technology, Vienna, Austria

¹eidenberger@tuwien.ac.at

²maia@prip.tuwien.ac.at

This contribution investigates the content-based feature extraction methods used in visual information retrieval, focusing on concepts that are employed for the semantic representation of media content. The background part describes the building blocks of feature extraction functions. Since numerous methods have been proposed we concentrate on the meta-concepts. The building blocks lead to a discussion of starting points for semantic enrichment of low-level features. The second part reviews features from the perspective of data quality. A case study on content-based MPEG-7 features illustrates the relativity of terms like “low-level,” “high-level” and “semantics”. For example, often more semantics mean just more redundancy. The final part sketches the application of features in retrieval scenarios. The results of a case study suggest that – from the retrieval perspective, too – “semantic enrichment of low-level features” is a partially questionable concept. The performance of classification-based retrieval, it seems, does hardly depend on the context of features.

1 Introduction

This contribution discusses the role of semantics in content-based image retrieval [12, 17]. It is organised in three sections. In the first section, we focus on *what (semantic) content-based features are*. Basic building blocks of signal processing-based low-level features are reviewed and starting points for enrichment of features are discussed and experimentally evaluated. The second section describes *what features extract*. It sketches analysis methods that allow for looking behind the scenes of feature extraction. Technically, the quality of feature extraction methods is judged from a quantitative point of view. Based on these insights, the third section

illustrates *how features are used*. It reviews application scenarios and – in a case study – compares the performance of content-based low-level features to context-free features in a typical retrieval scenario.

The paramount intention of this contribution is to give the reader a feeling for *how* content-based image features work and what enrichment with semantic information *means* (especially, on the data level). It is shown that, in practice, the borders between low-level feature extraction, semantic and context-free features almost vanish. For example, quantisation methods – frequently used in most image features – are nothing else but the introduction of domain knowledge into the feature extraction process. In this context, “domain” denotes concepts on a number of levels. It may stand for application domains (regular application case for semantic methods) but as well for technical domains (e.g. image properties). The essence of successful semantic image retrieval is to learn to handle the trade-off between such constraints and the benefit from narrowing the ambiguity of media content. Eventually, image retrieval is an ill-posed problem. For every gain in performance a considerable amount of generality has to be given away.

2 Building blocks of content-based image features

2.1 Signal Processing Building Blocks

In recent years, a remarkable number of content-based image features has been proposed, ranging from simple description of colour distributions to description of complex objects [12]. The visual part of the MPEG-7 standard [16] provides a comprehensive overview over the state of the art in low-level feature extraction. The building blocks of content-based features can be split into five groups (see Figure Fig. 1). Each building block provides enriched information that is further processed in consecutive data manipulation steps.

1. *Unitary time-to-frequency transformations* create a frequency representation of the media signal. Frequencies describe fundamental variations in the visual content. Low frequencies correspond to smooth changes (e.g. large uniform areas of background) and high frequencies correspond to abrupt changes (e.g. edge and corner information, but also noise). The Discrete Fourier Transformation (DFT) is the classic representative of such transformations. It is widely employed in audio retrieval. In visual retrieval, the Discrete Cosine Transformation (DCT,

e.g. compression) [19], Wavelet-based multi-resolution analysis [14] and the Angular Radial Transformation (ART) are more common. In the MPEG-7 standard, for example, DCT and ART are adopted in *Color Layout* descriptor and the *Region-Based Shape* descriptor, respectively [8].

2. *Parametric image transformations* map the input data to a condensed space. The mapping is controlled by parameters and usually not invertible. One example for parametric transformations is the Hough transformation [7]. It is employed to detect lines and to extract objects. The Radon transformation [21] is a second example for a parametric transformation. It is used to provide a rotation-invariant representation of image content. For this purpose, it is, for example, employed in MPEG-7 in the Homogeneous Texture descriptor that extracts characteristics of textures.
3. *Localisation methods* are used to select and describe data regions for further processing. State of the art techniques include the segmentation into rectangular macro-blocks, the definition of geometric primitives as regions of interest, spatial segmentation methods (e.g. by boundary detection, edge operators, etc.) and frequency filtering. The selection of the appropriate technique is mostly determined by the specific characteristics of the input data and the consecutive processing steps. MPEG-7 uses several localisation descriptors such as the Region Locator for regions of interest or the Contour-based Shape descriptor for spatial segmentation.

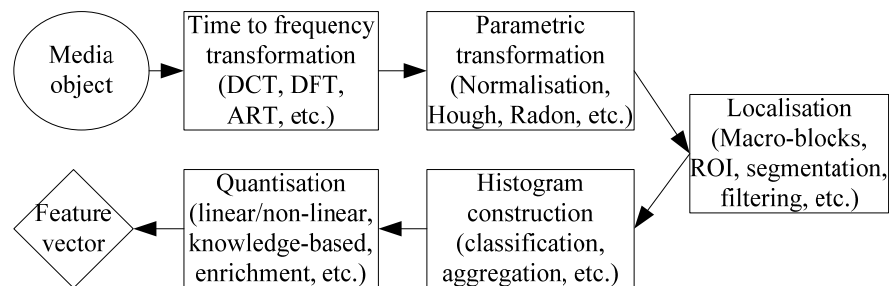


Fig. 1. Royal order of feature extraction building blocks.

4. *Histogram methods* are the methods of choice for data unification and conceptual understanding of types and frequencies of occurrence. They are conceptually similar to parametric transformations (e.g. they are not invertible). Based on (semantic) similarity criteria, data is

grouped (clustered) and counted. Such aggregation methods are not only applicable for colour properties and intensity distributions, but for any data population that contains a minimum of uniformity (e.g. edge types, frequencies, etc.). One application example in MPEG-7 is the Edge Histogram descriptor that provides information about the distribution of edge types in localised image regions.

5. *Quantisation methods* reduce and/or normalise the amount of high-frequency information in the feature data. Like aggregation, quantisation requires a certain amount of semantic knowledge. Quantisation is, for example, used for the reduction of storage space and for the satisfaction of limited bandwidth requirements. Methods include simple linear quantisation (e.g. normalisation of colour ranges by division into equally sized intervals), non-linear quantisation (e.g. variable intervals), and various types of “magic” quantisation. *Magic quantisation* puts an umbrella over all methods that consider additional (semantic) information such as domain rules, user information, etc. The bin-wise quantisation rules employed in the MPEG-7 *Scalable Color* descriptor are good examples for magic quantisation.

2.2 Starting Points for Semantic Enrichment

One of the most relevant ongoing activities in feature design is the semantic enrichment and interpretation of low-level features. Semantic enrichment endeavours to narrow the semantic gap. Since the perception of visual content is highly subjective, the same content may be interpreted differently by different users or even by the same user under different circumstances. Generally, human perception is based on three types of stimuli: generally perceived (not recognised) stimuli (e.g. colour or intensity distribution), specifically perceived (recognised) stimuli (e.g. object recognition), and pseudo-random stimuli (e.g. psychological, sociological, etc.). Retrieval of media objects exclusively by generally perceived properties is unsatisfactory. Mainly three sources of information are available for feature enhancement: information on the application domain (e.g. on media content), information on the user (e.g. user or retrieval preferences), and information on the characteristics of the specific features (e.g. statistical properties). Semantic feature research comprises the following (selected) groups of approaches:

1. *Feature data enrichment*. Additional knowledge can be induced (mainly in the aggregation and quantisation steps) by methods from statistics, artificial intelligence, neuronal networks, etc. For example,

domain knowledge on football could be used to identify field, ball and players from shape features (e.g. circularity). Application knowledge on the arts could be used to distinguish between classic and surreal paintings based on colour and edge features. Currently, many research efforts focus on the reduction of the search space by clustering [9]. Most clustering methods are unsupervised classification techniques (e.g. hierarchical (agglomerative or divisive) clustering [9], Self-Organizing Maps [11]). Such approaches divide the set of processed data (image features) into subsets (clusters) based on inherent data similarities (measured as metric distance). Hence, each cluster contains objects of more than average similarity.

2. *Context-based querying.* Integration of the media context in the feature extraction process is another starting point of semantic retrieval. Often, images are embedded in some type of document (e.g. web pages, multimedia presentations, text documents). In such situations, multimodal feature extraction, enrichment and retrieval based on semantic concepts may improve the relevance of media descriptions significantly. The joint application of text retrieval, content-based visual and audio retrieval can be performed by clustering, co-quantisation and during query execution.
3. *Feature knowledge modeling.* Eventually, though today semantic web technologies are almost exclusively employed for the annotation of markup text and for knowledge representation (e.g. RDF Schema, OWL, DAML+OIL), in the future they may also find applications in semantic enrichment of content-based media descriptions. Ontologies offer meta-concepts for the description of constraints and relationships among objects (for example, cardinality, domain and range restrictions, union/disjunction, inverse rules, etc.). In the context of content-based features ontology languages could be applied to specify constraints and relationships among descriptions, feature elements and semantic knowledge. Content-based ontology concepts would allow a clearer distinction between feature extraction models, descriptions and semantic metadata. Furthermore, they could provide the basis for rule-based feature derivation (e.g. localisation relationships, quantised colour histograms). Currently, such models are barely available. Many content-based features are a mixture of signal processing and the application (often, without reflection) of semantic knowledge.

3 Feature structure analysis

The first section has introduced how features are extracted from the media content and – in principle – how features can be lifted to a semantically higher level. This section investigates features from the perspective of data quality. Statistical data analysis offers tools that deserve more attention in visual information retrieval since these methods allow, for example, to judge whether an enrichment method causes more than just a higher level of information redundancy.

3.1 General-Purpose Feature Analysis Methods

Below, it is explained how quantitative data analysis methods can be employed to evaluate content-based media. It is shown how statistical analysis methods can be applied to judge various aspects of feature data quality. For example, cluster analysis methods and factor analysis methods (e.g. Principal Component Analysis) can be used to identify redundancies in the feature data. Topological clustering techniques are valuable tools to test the sensitivity of features for changes in the media data (e.g. noise, data loss). Statistical indicators can be applied to express the quality of feature data distribution and hence, the potential of features for similarity-based discrimination in retrieval.

The traditional evaluation scheme used in (visual) information retrieval employs recall and precision indicators computed for sample queries on well-known media collections. This process has its drawbacks and leaves the feature designer with a handful of open problems. The most relevant issue for application is defining a *ground truth* that reflects human similarity judgement appropriately independently of cultural aspects and other human peculiarities. Analysis of a feature requires embedding it in a querying framework and execution of hundreds of queries in order to guarantee statistical validity of the quality indicators. This process has to be repeated on every change in the feature transformation and it does not provide any hints on problems in the feature transformation. The bottom line is that recall and precision allow to estimate whether a feature performs well but not *why* it (which of its (semantic) properties) shows the observed behaviour.

9.3.1.1 The Big Picture of Quantitative Feature Analysis

The proposed supplementary evaluation procedure is a lightweight process that employs statistical data analysis methods (especially, Self-Organizing

Maps [11]) to evaluate information-theoretic quality aspects of feature transformations. A querying framework and ground truth information are not required. Quality indicators are derived directly from the feature data and allow conclusions on the statistical quality of the extraction process.

As for the traditional approach feature transformations are applied on predefined media collection. The resulting *feature data matrix* (feature vector elements by media objects) is normalised and investigated for characteristic properties (e.g. variance per variable), common properties and similarities by statistical methods. The essential point for understanding the key idea is that – in contrast to the retrieval situation – here, media objects are used to describe feature elements (e.g. colour histogram bins, shape moments). That is, technically, statistical methods are applied on the *transposed* feature data matrix. In statistical terms, the feature vector elements are the variables and the media objects are the cases. For example, in the experiments below MPEG-7 description elements (*Color Layout* coefficients, *Edge Histogram* bins, etc.) are fed – all described by the same set of media objects – into a Self-Organizing Map clustering algorithm to identify similarities between them.

Apart from applying statistical methods on visual features, another new aspect of the proposed evaluation process is that the feature in question is not just compared to itself but also to *reference data*. This reference data is provided by the content-based visual part of the MPEG-7 standard. Visual MPEG-7 descriptions are calculated for the predefined media collections. Comparing the feature vectors of the evaluated feature transformation to the reference data allows for gaining additional insights on the characteristics of a feature.

This proposed procedure has several advantages: Firstly, measurement is performed in a systematic way: one system (feature) is compared to another. Since the process is independent of the user (no user input required) it yields objective results. Secondly, results are application-independent. General data quality is measured instead of retrieval quality. Furthermore, no querying framework is required to apply this evaluation method. All necessary steps can be fulfilled with mathematical/statistical standard software (e.g. SPSS, Matlab).

3.1.2 Application Scenarios

A variety of questions including the following can be answered by statistical analysis:

- What is the "type" of a new feature? With respect to the MPEG-7 visual norm, is it a colour, texture, shape or motion feature or does it define en-

tirely new (semantic) criteria for visual media? Investigating proximities between a new feature and visual MPEG-7 descriptions may give valuable indications on promising starting points for closer (e.g. algorithm-based) examination of the new feature and help avoiding unwanted parallel developments.

- How robust is a new feature against rotation, scaling and other visual media transformations? Are the feature vector elements of original and transformed media data still similar after transformation? If not, do transformations change the characteristics of the feature transform?
- How robust is a new feature against noise? Do the characteristics of the feature vectors change if the media objects are noisy? This includes coding noise, i.e. artefacts introduced by lossy coding algorithms. Of course, sensitivity for every other type of noise can equally be tested, if the required media data is available.
- What is the effect of semantic enrichment on the data quality? Does an enriched feature represent new properties that are independent of those already identified by low-level features or does the semantic enrichment just cause higher data redundancy or make the feature extraction procedure more noise-prone?
- Does a feature mapping represent human visual similarity perception adequately? If the feature transformation is applied to two collections of similar media objects, are the corresponding feature vector elements similar, too?

Apparently, stronger answers can be given on these questions, if – in addition to recall- and precision-like evaluation procedures – statistical methods are considered.

3.1.3 Evaluation Workflow, Data Basis and Reference Data

The flow of work in the statistical evaluation procedure is as follows. Firstly, the new feature transformation is applied to predefined media collections. Numerical feature vectors are extracted. In the second step these feature vectors are merged with the pre-extracted MPEG-7 descriptions. After merging, data are normalised to a certain interval (e.g. $[0, 1]$) or particular moments (e.g. $\mu=0, \sigma=1$). Conventionally, all feature vectors together are addressed as the *feature matrix* (feature elements in columns, media objects in rows). On the feature matrix statistical operations are applied and indicators are derived. In the last step, these indicators are visualised and interpreted. Based on the interpretation the proposed feature transformation can be iteratively refined.

Various statistical methods exist that can be employed for evaluation. Principally, the three main areas relevant for visual information retrieval are univariate description, detection of similarities and detection of dependencies in the feature matrix (both multivariate). In earlier experiments we found these methods very useful [5]:

- Extraction of moments of first and second order of feature elements as well as computation of a discrete distribution of values for each element. The distribution reveals how often each value (down-sampled to a few bits) occurs. For example, it allows conclusions on how well a feature element utilises its data type.
- One- and two-dimensional cluster analysis of *feature elements* (not media objects!) for similarity assessment. *K*-means clustering and dendrograms for visualisation have proven to be sufficient in the one-dimensional case. Unfortunately, dendrograms become soon unreadable for larger numbers of elements. In this situation, two-dimensional clustering techniques (e.g. Self-Organizing Maps [11]) yield better results.
- Detection of dependencies of feature vector elements by factor analysis. Eigenvalues extracted from a data matrix (e.g. by a Principal Component Analysis) can be interpreted as hidden factors that have a linear influence on the data values. Elements (media properties) that are significantly influenced by the same factors (expressed by a factor loadings matrix) are obviously dependent of each other.

Principally, any media collection can be used for statistical analysis. The definition of a ground truth is not required. Still, it simplifies the interpretation process (especially, if semantic features are concerned), if the media collections have an inherent context. For some statistical evaluation methods only testbeds with a small number of media objects are manageable. Furthermore, statistical evaluation results depend to a certain degree on the a priori structure of the investigated media collection. Even though the media basis may be chosen arbitrarily, its structure has to be taken into account as a biasing factor in the interpretation of results (e.g. if cluster analysis is used)!

3.1.4 Summary

Semantically enriched features should have significantly different statistical properties than low-level features. Semantic enrichment desires to induce more meaning into the feature data. However, applying the proposed statistical analysis techniques reveals that often, this means just more redundancy. In simple words, semantically enriched features often look the same for different media content. The negative consequences of this effect for retrieval are obvious.

3.2 Case Study: Structure Evaluation of MPEG-7 Features

A few examples should make the suggested analysis approach and its advantages more transparent. Below, the statistical method introduced in the previous subsection are employed to judge the (semantic) quality of feature data extracted by selected content-based visual MPEG-7 descriptors. The analysis focuses on redundancies (e.g. from the application of semantic quantisation methods) and on sensitivity to content changes and noise.

3.2.1 Case Study Setup

We analyse the majority of the content-based visual MPEG-7 descriptors. All colour descriptors: *Color Layout*, *Color Structure*, *Dominant Color*, *Scalable Color*, all texture descriptors: *Edge Histogram*, *Homogeneous Texture*, *Texture Browsing* and one shape descriptor: *Region-based Shape*. The other basic shape descriptor, *Contour-based Shape*, is not used, because it produces structurally different descriptions that cannot be transformed to data vectors measuring on interval scale. Description extraction is performed employing the MPEG-7 experimentation model (XM, [23]) of MPEG-7 Part 6: Reference Software. In the extraction process each descriptor is applied on the entire content of every media object.

The descriptors are applied on three image collections: the Brodatz dataset [2] (112 monochrome images, 512x512 pixel), a subset of the Corel dataset [24] (260 colour photos, 460x300 pixel, portrait and landscape) and a dataset with coats-of-arms images [1] (426 synthetic images, 200x200 pixel). The evaluation is performed in the following steps: description extraction, normalisation, extraction of statistical indicators, quantisation and extraction of distributions, hierarchical cluster analysis, computation of topological cluster structures and factor analysis. After the description extraction, the resulting XML-descriptions are transformed into a data matrix with 798 lines (media objects) and 314 columns (description elements).

Mean and standard deviation are used as primary indicators for description elements. To identify the distribution of values of description elements over N media samples, the coefficients of the data matrix are quantised to ten bins. For the hierarchical cluster analysis a single-linkage algorithm with squared Euclidean distance measurement is used. The results are depicted as dendrograms on a relative scale from 0 (identical) to 25 (not similar). Self-Organizing Maps (SOMs) [11] are employed for topological clustering. SOMs are calculated with a hexagonal layout (every non-border cluster has six neighbours). For cluster adaptation, a Gaussian neighbourhood kernel is employed. Maps are initialised randomly. For factor extraction a Principal Component Analysis (analysis of the coefficients of the

correlation matrix) is used [13]. All Eigenvalues greater than one are selected as factors. To simplify the interpretation process, a Varimax rotation is performed on the factor loadings matrix. Factor analysis can only be applied on elements with existing variance. For the Brodatz dataset 225 description elements fulfil this requirement, for the Corel dataset 311 and for the coats-of-arms dataset 310. For the remaining elements, the description extraction algorithms comes up with exactly the same values independent of the analysed content.

3.2.2 Redundancy Analysis

In this analysis we try to identify whether the description elements extracted from visual content are *unique* or not. Redundancy information is highly valuable for two major reasons. It may influence how descriptions are organised in description schemes (*efficiency of application*). It is obviously not desired to combine certain descriptors to a description scheme (e.g. as a means of semantic enrichment) if it is well known that the descriptors are highly redundant for the concerned media class. Additionally, it can be used as a supplementary method to the MPEG-7 binary format [16] for compression of descriptions (e.g. for specific classes of content). This helps to further reduce the amount of space and bandwidth needed in visual retrieval systems (*efficiency of representation*).

A first striking result revealed by the hierarchical cluster analysis is the high self-similarity of the elements of the *Homogeneous Texture* descriptor for any type of media (see Table 1). For the Brodatz dataset (rich textures) and the coats-of-arms dataset (poor textures) all description elements form a single cluster with a maximum distance of 4%. Interestingly, the *Edge Histogram* descriptor forms five to ten clusters with ten to 15 elements for any type of content. The elements of these clusters are self-similar but the distance between the clusters is relatively large.

Table 1. Results of hierarchical cluster analysis: number of clusters and distances between clusters. The maximum distance is given in percent (where 100% would be the distance of a vector of "0" values to a vector of "1" values).

<i>Descriptor</i>	<i>Media collection</i>	<i>No. of clusters</i>	<i>Maximum distance between clusters</i>
<i>Homogeneous Texture</i>	Brodatz,	1	4%
	Coats-of-arms		
	Corel	2	20%
<i>Edge Histogram</i> other	any	5-10	12%-20%
	any	>5	>20%

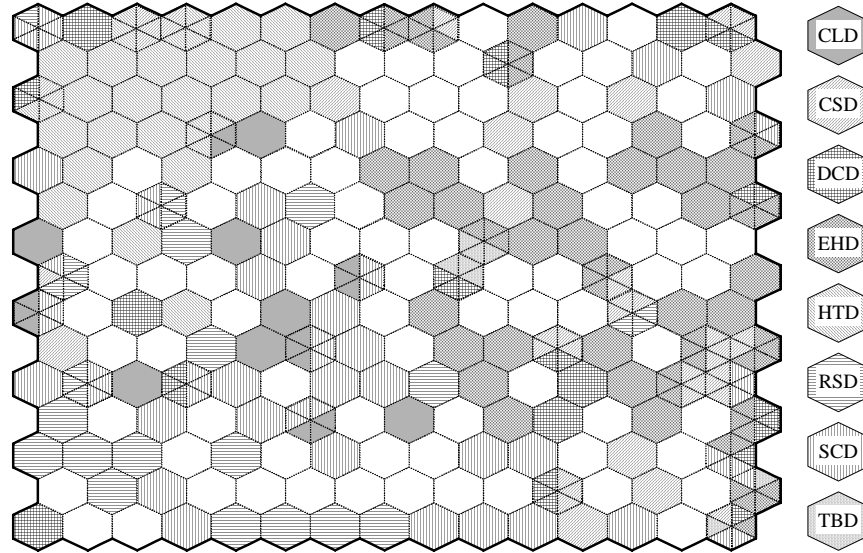


Fig. 2. Self-Organizing Map of MPEG-7 description elements for the coats-of-arms dataset. Neighbouring clusters contain similar description elements. Since every non-border cluster has six neighbours, clusters are shown as hexagons. Cluster populations are depicted as textures (CLD: *Color Layout*, CSD: *Color Structure*, DCD: *Dominant Color*, EHD: *Edge Histogram*, HTD: *Homogeneous Texture*, RSD: *Region Shape*, SCD: *Scalable Color*, TBD: *Texture Browsing*). If clusters are shared between descriptors, hexagons are split into triangular regions.

Analysing the SOMs for the three media collections (see Figure Fig. 2) supports the first impression of the hierarchical analysis. *Homogeneous Texture* lays a fine-meshed net over the investigated media property. The *Edge Histogram* descriptor forms clusters that contain slightly more elements. The net of the *Edge Histogram* is wide-meshed but the descriptor covers a larger area of the variance in the media data. All other descriptors form rather small 2d clusters for any type of content.

A more detailed view can be obtained from the factors extracted by factor analysis algorithms (see Table 2). For the Brodatz dataset 34 factors explain 225 description elements (the remaining elements have zero variance). 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 more details and, generally, descriptions should be less redundant on material with richer content. 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 identical colour component val-

ues. Maybe certain characteristics (e.g. quantisation) in the extraction algorithm implemented in the MPEG-7 XM cause this phenomenon.

Table 2. Results of factor analysis: number of extracted factors and explained variance. Only elements with existing variance are considered.

<i>Media collection</i>	<i>Elements with existing variance</i>	<i>Factors</i>	<i>Explained variance (all)</i>	<i>Explained variance (first factor)</i>	<i>Redundancy relationship</i>
Brodatz	225	34	89%	15%	7:1
Corel	311	69	85%	12%	9:2
Coats-of-arms	310	71	80%	6.7%	9:2

Several observations can be made from these analysis results. Generally, the MPEG-7 descriptors generate results of high redundancy. The *magic quantisation by feature characteristics* used in most descriptors may be one explanation for this observation. Especially, all MPEG-7 descriptors are highly redundant for monochrome media content. On the other hand, all bins of *Color Layout* are highly un-similar for any type of media content and independent from all other elements. Especially the first element (luminance DC coefficient) seems to be a good indicator for global shape information even for complex scenes (as *Region-based Shape* should be). The elements of the *Homogeneous Texture* descriptor are – independent of the media – highly self-similar and redundant. The ideal – content-independent – description scheme for visual content seems to be *Color Layout* (because of the first element), *Dominant Color*, *Edge Histogram* and *Texture Browsing*. This DS provides a maximum of semantic context (on the MPEG-7 level) at a minimum of redundancy.

3.2.3 Sensitivity Analysis

This analysis tries to give indication on the sensitivity of the descriptors on varying media content. In detail, three forms of sensitivity are investigated: firstly, sensitivity of colour descriptors for monochrome content, secondly, sensitivity of colour descriptors for content with few colour shades (e.g. animations) and finally, sensitivity of the texture descriptors and *Region-based Shape* for coarse, medium and fine structures in the content. Ideally, the descriptors should provide surjective mappings from the visual content to the feature space. These mappings should be robust against variations in the quality of the content (e.g. presence of colour information, resolution). Analysing the sensitivity allows to judge to which extent "bad" (e.g. bleached) input affects the quality of the descriptions.

The main indicators for sensitivity are mean and standard deviation. For a uniformly distributed element on the interval [0, 1] with a mean of 0.5, the maximum standard deviation is 0.346. In the evaluation the standard deviation should be 0.2 or higher (using at least an interval of 40% of the data range for 66% of all media objects) to be acceptable. Then, the description element can be considered as being sufficiently discriminant to distinguish media objects independently of variations in the content.

Table 3 summarises the average means and standard deviations of the colour description elements. *Color Layout* performs badly on monochrome data (Brodatz dataset). Only six of twelve bins have a standard deviation greater than zero: the DC and AC coefficients of the luminance channel. Whenever colour is present – independent of the number of gradations – *Color Structure* performs excellently. The average standard deviation is 0.25. Therefore, the element values are distributed over the entire range of possible values. For monochrome content, *Scalable Color* is not able to derive meaningful descriptions. For the Corel dataset, *Scalable Color* results are excellent. The average standard deviation is 0.3. *Edge Histogram* performs excellently on any type of media (see Table 4 for details on texture and shape descriptors). The *Homogeneous Texture* descriptor performs poorly on colour images, especially if they have few colour shades and textures in them. Finally, the *Region-based Shape* descriptor measures excellently on any type of media. These findings are supported by the cluster analysis results. Most clusters are on distance level lower than 20%. Hardly any clusters exist at average distance (20% to 60%). Cluster structure and clusters size varies widely for different content.

Table 3. Average mean and standard deviation of colour description elements. Only elements with existing variance are considered in the averaging process.

<i>Descriptor</i>	<i>Media collection</i>	<i>Average mean</i>	<i>Average standard deviation</i>
1. <i>Color Layout</i>	Brodatz	0.7	0.1
	Corel	0.55	0.2
	Coats-of-arms	0.65	0.15-0.2
<i>Color Structure</i>	Brodatz	0.85-0.9	0.05-0.15
	Corel,	0.5	0.25
	Coats-of-arms		
<i>Dominant Color</i>	any	0.45-0.5	0.3
<i>Scalable Color</i>	Brodatz	0.4	0.3
	Corel	0.5	0.3
	Coats-of-arms	0.4-0.5	0.15

Table 4. Average mean and standard deviation of texture and shape description elements.

<i>Descriptor</i>	<i>Media collection</i>	<i>Average mean</i>	<i>Average standard deviation</i>
<i>Edge Histogram</i>	any	0.5	0.25-0.3
	<i>Homogeneous Texture</i>	Brodatz	0.65-0.7
<i>Texture Browsing</i>	Corel	0.75	0.1
	Coats-of-arms	0.75	0.05
	Brodatz,	0.2-0.3	0.2-0.25
<i>Region-based Shape</i>	Corel		
	Coats-of-arms	0.1	0.05
	any	0.5	0.2-0.25

3.2.4 Summary

All colour descriptors work excellently on photos (high-frequency input) but *Color Layout*, *Color Structure* and *Scalable Color* perform badly on artificial media objects with few colour gradations, and very badly on monochrome content. *Edge Histogram* is by far the best texture descriptor (high sensitivity, low redundancy). *Homogeneous Texture* is highly sensitive to the analysed media content and the variance of results is small. *Region-based Shape* is a good descriptor that can be applied to any type of media content.

These results are partially surprising. MPEG-7 descriptions are highly redundant, sensitive to noise and provide only little ground for discrimination. Especially, the quantisation methods proposed in MPEG-7 (semantic enrichment based on feature knowledge) cause a regrettable loss in the quality of media descriptions. This behaviour is a striking example for the dangers of semantic enrichment. Descriptions may become less discriminant and drift to represent superficial media properties. One possible cure for such structural problems would be to revise the application of additional knowledge in concerned descriptors (e.g. the quantisation steps suggested in the MPEG-7 standard).

4 Features at work

4.1 (Semantic) Feature-Based Retrieval Approaches

Multimedia information retrieval systems are often distinguished by the querying paradigm. Frequently used methods are querying by example, querying by sketch, iconic querying, and querying by example groups. Though being important for the user, the selected querying paradigm does not influence the retrieval process. Media comparison is always based on descriptions (feature vectors). If not available beforehand, descriptions have to be extracted during the querying process (e.g. of sketches or example images).

In most systems, one of two retrieval processes is employed: retrieval based on the *vector space model*, or retrieval based on *probabilistic inference*. The vector space model assumes that media descriptions are points in a vector space and that this vector space has a geometry (mostly, Euclidean geometry is assumed) [6]. Then, unsimilarity of media objects can be measured as metric distance of media descriptions (e.g. City Block distance, Euclidian distance, Minkowski distances). The vector space model is successfully used in text information retrieval. Unfortunately, applied to MMIR, two problems arise. Firstly, it is not clear what type of geometry (distance function) fits to human similarity perception. Secondly, often differently extracted media descriptions require different distance measures. The selection of features for retrieval and the usage of multiple distance measures are non-trivial, still open research problems (see, for example, [4]).

Probabilistic inference models use media descriptions and a priori probabilities (computed from statistics based on e.g. human relevance information) to compute differentiated a posteriori probabilities that can be used for retrieval [6]. Employed models are mostly based on Bayesian networks, i.e. topologies that represent dependencies among features. The major advantage of probabilistic inference models over the vector space model is that they avoid the problem of explicitly defining similarity measures. The main disadvantages are that sample data are required and that fast-learning relevance feedback algorithms (see below) are hard to define. Furthermore, the *independence constraint* assumed in many inference models (e.g. the Binary Independence Retrieval Model [6]) may be too restrictive for real-world scenarios. Human perception is highly sensitive and culturally loaded. For example, a picture of two adults with a child

may be perceived differently from the same picture without the child (family vs. lovers).

Severe problems as the semantic gap and polysemy have lead to the insight that visual retrieval may be modelled best as an iterative process. Retrieval steps should be directed by the user's relevance feedback. Today, one refinement technology outperforms most other approaches: classification by Support Vector Machines (SVM) [18]. SVMs separate a given set of binary labelled data (relevant/irrelevant) by their maximum margin, i.e. a hyper-plane in maximum distance of the two groups. Moreover, kernel functions are employed to project the input data to a higher-dimensional (less densely populated) feature space. This transformation simplifies the separation process. The advantages of SVMs are straightforward application and high performance. SVMs are easy to apply, since only two groups (relevant and irrelevant media objects) have to be distinguished by the user. The outstanding performance of SVMs may at the same time be their major weakness. SVMs are prone to overfitting. Hence, careful selection of training data is a crucial step in SVM application.

It should be noted that none of these retrieval approaches makes assumptions on the *semantic level* of the employed media metadata. As long as certain mathematical requirements are fulfilled (e.g. measurement on interval scales), the feature data may be of arbitrary shape. In experimental evaluation it may turn out that semantic concepts and data quality considerations are eventually irrelevant for retrieval application. This suspicion may especially be true in cyclic retrieval scenarios based on classification. The following case study evaluates this question.

4.2 Case Study: CB Features vs. Semantic Features

In this section we turn the considerations from the previous sections to the extreme and regard context-free (random) features as semantic features, i.e. as expressions of visual content that are human-like, hence (almost) unpredictable for machines. In a retrieval scenario, we compare the performance of the context-free image features to sophisticated signal processing-based features (the MPEG-7 image features [16]). Our context-free features show a maximum of the data anomalies investigated in the evaluation section. However, they are employed in the same retrieval processes as the MPEG-7 descriptions.

4.2.1 Case Study Setup

The experiments are undertaken in an environment that simulates user-centred visual information retrieval. Querying is performed as cyclic refinement of results by relevance feedback. That is, retrieval results are provided by classification. Kernel-based Support Vector Machines [25] are employed for active learning of classifiers. In the automated evaluation process the training samples are taken from a pre-defined ground truth (representing the visual similarity judgement of users). For the sake of realistic results only few examples are used for training. The trained classifiers are employed to discriminate the populations of queried collections into groups of relevant and irrelevant media objects.

The extraction of MPEG-7 features is performed by using the MPEG-7 experimentation model. The context-free features are extracted from the test dataset using the Matlab algorithms discussed in the results section. SVM-Light [10] is employed for SVM classification as well as for the computation of recall and precision values.

Training is performed on the fraction of the data (random selection of vectors) given in the results section (1%, 10% etc.). Classification is always performed on the entire ground truth group of positive examples and an equal number of negative examples. This rule allows for easier interpretation of results, since every feature below 50% recall/precision can be discarded for being dominated by guessing.

The following content-based visual MPEG-7 descriptors are employed in the analysis process [16]: *Color Layout*, *Color Structure*, *Dominant Color*, *Edge Histogram*, *Homogeneous Texture*, *Region-based Shape* and *Scalable Color*. In total, every media object is described by 306 feature elements. Two collections are merged to form the test media dataset: parts of the well-known Corel dataset (1615 colour images) and the UCID dataset provided by the University of Nottingham [22] (1338 colour images). The total 2953 images are split into 17 groups of conceptually similar content. Noteworthy, the largest ground truth groups contain eight times more images than the smallest groups.

4.2.2 Experimental Results and Interpretation

Algorithm 1 computes a simple context-free feature for the media objects in the test dataset. 300 uniformly distributed random numbers per media object are employed as features. The dimensionality has been chosen to be comparable in length to the MPEG-7 features. Table 5 (1% training data) lists the performance of the semantic/random feature in relation to MPEG-7.

```
function extractSemanticFeature
```

```

inHandle = fopen('mediaFiles.dat');
data=zeros(1:1);
i=1;
while (feof(inHandle)==0)
    fname = fgetl(inHandle)
    for j=1:300
        data(i, j) = rand;
    end
    i=i+1;
end
fclose(inHandle);

```

Alg. 1. Extraction of simple random features.**Table 5.** Performance of Alg. 1 in comparison to MPEG-7 (1% training data).

<i>Feature</i>	<i>Kernel</i>	<i>Recall</i>		<i>Precision</i>	
		<i>%</i>	<i>Rank</i>	<i>%</i>	<i>Rank</i>
MPEG-7	Linear	77.55	2	57.91	3
Random (A1)	Poly	74.61	4	51.02	9
MPEG-7	Poly	74.41	5	59.58	1
MPEG-7	Radial	73.76	6	59.21	2
Random (A1)	Linear	64.82	8	51.20	8
Random (A1)	Radial	62.44	9	51.52	7

```

function extractSemanticFeature
% [...] see Algorithm 1
for j=1:300
    data(i, j) = i/numberOfFiles*rand;
end
% [...] see Algorithm 1

```

Alg. 2. Extraction of improved random features.

The random feature performs inferior to MPEG-7 if sufficient training data is available. For 10% training data (data table not given here), recall is 11% and precision is 17% behind. Still, in situations where just a handful of training samples are available, the gap to MPEG-7 shrinks frighteningly. For the best kernel and 1% training data the random feature is only 3% recall behind the MPEG-7 features. These findings suggest that in SVM-based content-based features (semantic or not) are just a ground for discrimination. However, the more training samples are available, the larger the gap to the top performing features becomes. In terms of precision the random feature is always very close to the guessing border (50%). It seems that the random feature does not allow for efficient discrimination.

Our suspicion is that this deficit is caused by the uniform distribution of random numbers used (high redundancy!). In average, all random feature vectors are relatively similar.

Hence, in a verification experiment an improved random feature is employed to compete with MPEG-7. In equation 1, the feature f_i of the i th media object is a vector of j columns. N is the number of media objects in the queried collection. $random()$ returns a uniformly distributed random number. This feature associates media objects with random numbers of significantly different variances (frequencies). Alg. 2 shows the Matlab formulation of the feature extraction algorithm. Since Eq. 1 combines the random function with a step function, the feature is called *random step* in the results in Table 6.

$$f_i = \left\langle \frac{i}{N} random() \right\rangle_j \quad (1)$$

The averaged performance indicators show that the random step feature performs astonishingly well. For 1% training data it outperforms MPEG-7 in terms of recall by 10%. Even its precision is competitive with just 1.5% behind MPEG-7. If 10% feature vectors are available for training (data table not given here), the difference in recall is still 9% and precision is head to head to MPEG-7. Astonishingly, random step matches the precision of MPEG-7. Since it is superior in terms of recall, the surprising conclusion of our experiment is that this simple “semantic” feature fulfils its purpose just as content-based MPEG-7 descriptions do.

Table 6. Performance of Alg. 2 in comparison to MPEG-7 (1% training data).

<i>Feature</i>	<i>Kernel</i>	<i>Recall</i>		<i>Precision</i>	
		<i>%</i>	<i>Rank</i>	<i>%</i>	<i>Rank</i>
Random Step (A2)	Poly	87.40	2	58.09	5
MPEG-7	Linear	77.55	3	57.91	6
Random Step (A2)	Linear	77.44	4	66.42	1
Random Step (A2)	Radial	76.80	5	61.70	2
MPEG-7	Poly	74.41	7	59.58	3
MPEG-7	Radial	73.76	8	59.21	4

From what we have learned in these experiments we advocate to shift the feature-related expectations of user-centred content-based retrieval. The role of media features is mainly to provide a *ground for discrimination* of user-labelled (similar and unsimilar) media objects. It is only of minor relevance if this ground is to 100% derived from media content. (This is a silent assumption in semantic enrichment, anyway.) Therefore, from our

perspective it would be preferable to speak of *image identification strings* instead of (semantic) media descriptions or features.

The following characteristics of content-based visual retrieval may be one major reason why SVM classification does not require content-based features. Human similarity judgement is goal-centred, influenced by many complex factors and therefore almost unpredictable. Hence, it is a very difficult task to represent it adequately by content-based features. For general application, it may be a more reasonable approach to abandon extracting content-based features in favour of maximising the discrimination ability of features. Eventually, content-based visual retrieval shifts from signal processing more and more towards information retrieval. The classifier used becomes the core component of the user-centred retrieval process. Currently, SVMs are the tool of choice for the imitation of human visual similarity perception.

4.2.3 Application of Context-Free Features

Without doubt, these findings make the application of sophisticated signal processing and semantic enrichment in user-centred retrieval partially questionable. However, the positive consequences of querying by classification of image ID strings are extraordinary. Image ID strings can be computed (assigned) quickly and do not have to be stored in a repository. Retrieval becomes a light-weight process. Systems can be developed and debugged significantly easier, since an entire level of sophistication falls away. Eventually, the results for the best context-free features are (despite higher redundancy and other data anomalies) even better than the results for content-based features. Context-free image ID strings seem to provide more space for discrimination (the desired type of semantics). This is a serious advantage in a research domain where it is silently accepted that the human influence (visual similarity judgement) may appear in an almost arbitrary fashion.

Furthermore, image ID strings allow for real-time feature extraction. Actually, it is sufficient to assign media objects with image ID strings from a pre-defined catalogue. Even this association has to be performed only once on the first retrieval iteration. The resulting retrieval process is light-weighted and can be performed ad hoc. All sophistication is encapsulated in the classifier. Hence, a retrieval situation is fully described by the Lagrange coefficients of the trained SVM. For frequently appearing queries it could make sense to store these coefficients as a *semantic feature* (the additional knowledge is provided by the user's relevance judgement). Moreover, features that can be constructed in real-time allow for the implementation of flexible multimedia database management technologies such as

the mediator paradigm [20]. If content similarity is sufficiently described by human feedback encapsulated in straightforward classifiers, it can be handled human-like flexibly.

However, some open problems remain. Content-based visual retrieval is often implemented as a two-step process. In the first step, distance measures and content-based features are used to retrieve a first result set. This set is iteratively refined by relevance feedback in the second step. Obviously, image ID strings cannot be employed for distance-based querying in metric spaces (e.g. by following the vector space model). Therefore, we suggest to replace the first step by a randomly chosen selection of media objects. If available, domain knowledge can be used as well.

4.2.4 Summary

SVM-based classification of users' relevance judgement is – especially on smaller collections – a hard to beat retrieval method. In fact, SVMs perform so well, it hardly matters *what* is classified. From the experiments in this case study we can see that simple context-free features perform as well (or even better) as sophisticated signal processing operations, for example, the content-based visual MPEG-7 features. SVMs require features only as an adequate *ground for discrimination*. Hence in this context, we consider it more precise to speak of image ID strings instead of features. The best image ID string identified in the experiments is random-based. However, from the insights of the data analysis section we understand that a uniform random function would not provide enough space for discrimination. Therefore, the random function is augmented by a simple step function. This construction outperforms the MPEG-7 features. These surprising findings show the *crucial power of the retrieval step*. Well-discriminating features can be computed in real-time during the retrieval process. Image retrieval turns into a more flexible process that is exclusively based on human visual similarity perception.

5 Conclusions

This contribution intends to:

- make the meaning and structure of content-based features transparent and explain what semantic enrichment means for the structure and content of visual features.
- explain in simple words the frequently used methods for feature extraction, semantic enrichment and cyclic retrieval.

- make the reader familiar with techniques for statistical analysis of feature data and explain what can be expected from content-based features.
- uncover the partial contradiction in the requirements of content-based features that should, at the same time, summarise the media content adequately and provide an efficient ground for discrimination.

One major conclusion of this contribution is that modern classification methods are such powerful retrieval methods; they can classify basically *any* kind of data efficiently. We believe that – compared to the users' relevance input (on media objects) – the shapes of the data vectors employed for media description (randomly chosen, semantically enriched, etc.) and their data quality are only of minor importance for the quality of retrieval results.

References

1. Breiteneder C, Eidenberger H (1999) Content-based image retrieval of coats of arms. Proc. of IEEE Multimedia Signal Processing Workshop, pp 91-96
2. Del Bimbo A (1999) Visual information retrieval. Morgan Kaufmann
3. Chang SF, Sikora T, Puri A (2001) Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology 11/6: 688-695
4. Eidenberger H, Breiteneder C (2002) Macro-Level Similarity Measurement in VizIR. Proc. of IEEE ICME, Lausanne, Switzerland, pp 721-724
5. Eidenberger H (2004) Statistical analysis of the MPEG-7 image descriptors. ACM Springer Multimedia Systems Journal 10/2: 84-97
6. Fuhr N (2001) Information Retrieval Methods for Multimedia Objects. In: Veltkamp RC, Burkhardt H, Kriegel HP (eds) State-of-the-Art in Content-Based Image and Video Retrieval. Kluwer, Boston, pp 191-212
7. Hough PVC (1962) A Method and Means for Recognizing Complex Patterns. US Patent 3,069,654
8. International Standards Organization (2002) MPEG-7 Information Technology – Multimedia Content Description Interface – Part 3: Visual. ISO/IEC 15938-3:2002(E)
9. Jain AK, Murty MN, Flynn PJ (1999) Data Clustering: A Review. ACM Computing Surveys 31/3: 264-323
10. Joachims T (last visited 2007-08-01) SVM light. svmlight.joachims.org
11. Kohonen T (1990) The self-organizing map. IEEE Proc 78/9: 1464-1480
12. Lew MS (2001) Principles of visual information retrieval. Springer, Berlin
13. Loehlin JC (1998) Latent variable models: An introduction to factor, path, and structural analysis. Lawrence Erlbaum Assoc, Mahwah, NJ
14. Mallat SG (1989) A Theory of Multi-Resolution Signal Decomposition: The Wavelet Representation. IEEE PAMI 11: 674-693
15. Manjunath BS, Ohm JR, Vasudevan VV, Yamada A (2001) Color and texture descriptors. IEEE CSVT 11/6: 703-715

16. Manjunath BS, Salembier P, Sikora T (2002) Introduction to MPEG-7. Wiley
17. Marques O, Furht B (2002) CB image and video retrieval. Kluwer, Boston
18. Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B (2001) An Introduction to Kernel-based Learning Algorithms. IEEE TANN 12/2: 181-202
19. Rao KR, Yip P (1990) Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, Boston
20. Santini S, Gupta A (2004) Mediating imaging data in a distributed system. Proc. of SPIE Storage and Retrieval Methods and Applications for Multimedia Conference, San Jose, CA, pp 365-376
21. Sanz JLC, Hinkle EB, Jain AK (1998) Radon and Projection Transform-Based Computer Vision. Springer, Berlin
22. Schäfer G, Stich M (2004) UCID - An uncompressed colour image database. Proc. of SPIE Storage and Retrieval Methods and Applications for Multimedia, Conference San Jose, CA, pp 472-480
23. TU Munich (last visited 2007-08-01) MPEG-7 experimentation model. www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/e_mpeg7.html
24. University of California Berkeley (last visited 2007-08-01) Corel dataset website. elib.cs.berkeley.edu/photos/corel/
25. Vapnik VN (1995) The nature of statistical learning theory. Springer, Berlin