A new perspective on visual information retrieval

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

ABSTRACT

Visual information retrieval (VIR) is a research area with more than 300 scientific publications every year. Technological progress lets surveys become out of date within a short duration. This paper intends to shortly describe selected important advances in VIR in recent years and point out promising directions for future research. A software architecture for visual media handling is proposed that allows handling image and video content equally. This allows to integrate both types of media in a singe system. The major advances in feature design are sketched and new methods for semantic enrichment are proposed. Guidelines are formulated for further development of feature extraction methods. The most relevant retrieval processes are described and an interactive method for visual mining is suggested that really puts "the human in the loop". For evaluation, the classic recall- and precision-based approach is discussed as well as a new procedure based on MPEG-7 and statistical data analysis. Finally, an "ideal" architecture for VIR systems is outlined. The selection of VIR topics is subjective and represents the author's point of view. The intention is to provide a short but substantial introduction to the field of VIR.

Keywords: Visual Information Retrieval, Content-based Image Retrieval, Content-based Video Retrieval, Survey, Media Representation, Feature Extraction, Similarity Definition, Evaluation

1. INTRODUCTION

This is a paper on retrieval of visual objects (images and videos) by content. In the year 2003 it is probably one of more than thousand papers in this area of research. In 2002 the IEEE alone has published more than 700 retrieval papers. Figure 1 depicts the increase in visual retrieval publications since 1981 (on basis of the IEEE digital library). Due to the increase of cheaply available (digital) image and video cameras and the increasing power of affordable computer systems visual information retrieval becomes more and more popular as a research discipline. Since 1994 more than hundred papers (=new ideas?) have been published every year.

In this paper we try to fence off important areas of visual information retrieval (VIR). For each area we will shortly describe important past advances and point out relevant, currently ongoing activities. The main focus of the paper is on arguing for new perspectives on selected VIR problem areas. In our opinion, the basic building blocks of VIR research are media management, feature design, querying, evaluation and system design. Each of these areas will be discussed in one section.

Our motivation is that, even though significant advances have been achieved and, by now, a large number of freely available mature VIR systems exists, VIR techniques are not adopted to an adequate extent in relevant application domains (e.g. digital libraries). One major reason may be the discrepancy of hopes associated with VIR (querying by *semantic* similarity) and the reality implemented in most prototypes (querying by low-level features). For example, it is annoying trying to retrieve Hollywood kisses in a movie database by colour, texture and shape features. On the technical level this fact is called "semantic gap"[19].

Even though in recent years a large number of approaches have been proposed to close – or at least narrow – the semantic gap (e.g. semantic enrichment of features, kernel-based learning to find relevant media objects) the potential of VIR still seems to be judged from the performance of the classic prototype systems. Clarifying the state of the art as well as future potentials is certainly an important task if VIR should have a future as a *practically* relevant addition to existing media management and retrieval tools (based on text). From the author's experience, one point should be stressed as most important: VIR technology is able to fulfil sophisticated semantic retrieval tasks but it is *not* able to replace human perception.

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

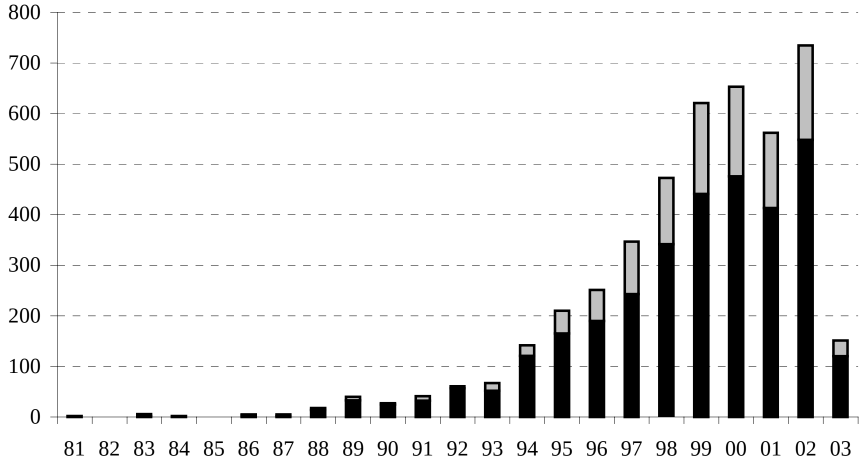

Figure 1: Number of papers in IEEE digital library containing "image retrieval" (black) or "video retrieval" (grey) in bibliographic data. (year 2003: status of 1st October 2003).

This paper reviews VIR from a subjective point of view: It reflects the author's opinion. The organisation is as follows. Section 2 points out relevant related work. The basic VIR building blocks are discussed in consecutive sections: Section 3 visual media, Section 4 visual feature design, Section 5 the retrieval process (similarity definition, interaction), Section 6 evaluation and, finally, Section 7 aspects of VIR system design.

2. BACKGROUND: VIR STATE OF THE ART REPORTS

A handful of VIR publications exists that survey the state of the art. Most of them reflect in organisation and content the perception of VIR of the time when they were written. Below, firstly, we will name a few outstanding representatives and try to sketch their view of VIR. The section will be concluded with remarks on ongoing activities to summarise recent findings in this area of research.

In the book "Image and Video Processing in Multimedia Systems"[14] by Furht, Smoliar and Zhang the state of the art of VIR up to the publication date (1996) is described. The authors start with a system model of content-based image retrieval (CBIR), describe image features (distinguished classically in colour, texture and shape features) and video features (shot detection and camera operation detection), indexing approaches for high-dimensional feature vectors, methods for interactive querying and evaluation based on ground truth information and retrieval quality indicators (recall and precision). Additionally, promising application domains are described and case studies for video visualisation are given.

"Image Retrieval: Past, Present and Future"[18] by Rui, Huang and Chang (1997) concentrates on CBIR. Again, the organisation is classic. Features are split into colour, texture and shape and high-dimensional indexing as well as dimension reduction (e.g. by principal component analysis) are important topics. Well-known CBIR prototype systems (QBIC, Virage, Retrievalware, Photobook, VisualSEEk, MARS) are described in detail. Additionally, this paper was the first survey that described Gabor wavelets as the best suited (in terms of perception) for time to frequency transformation. It led the way for future research as it stressed the importance of putting the "human in the loop" of interactive querying (relevance feedback) and of semantic enrichment of low-level features by artificial intelligence methods. Also, it stated the evident demand for benchmarking initiatives for CBIR systems and gave a first outlook on the MPEG-7 project.

The book "Visual Information Retrieval"[2] by Del Bimbo (published in 1999) is organised by feature groups. As in all other VIR surveys up to now, image and video retrieval are treated separately. For each group of features (colour, texture, shape, motion (shot segmentation only)) extraction methods, distance measures and application examples are described. Classic topics like indexing, evaluation and system design are briefly described. To the author's knowledge this book introduces the terms "semantic gap" and "multi-resolution analysis" for the first time in a survey. The hypothesis of multi-resolution analysis is that using iteratively computed 2D wavelet coefficient matrices as features is sufficient for retrieval. Additionally, the author describes in detail the usage of image features in (spatial) combinations.

The journal paper "Content-based Image Retrieval at the End of the early Years"[22] by Smeulders, Worring, Santini, Gupta and Jain (2000) gives a broad view on CBIR. For the first time selected features are not described in detail but the characteristics of features classes (mainly shape features) are abstracted. Similarity measurement is treated as a topic independently of feature extraction, and distance measures and their geometric foundations are discussed in detail. The importance of learning methods for iterative query optimisation is stressed. Additionally, system aspects (indexing, evaluation, etc.) and techniques of related fields (e.g. edge detection, shape description) are discussed.

Finally, "Content-based Image and Video Retrieval"[16] by Marques and Furht gives only a short overview over the various building blocks of VIR systems and concentrates on conservative techniques. Its major strength lies in the description of a vast number of prototypes for both image and video retrieval. Additionally, design issues of image and video retrieval systems are discussed and case studies are given.

Since hundreds of new ideas are introduced in VIR every year, every survey can only stay up to date for a very short duration. Among the recent publications, the papers on the visual MPEG-7 descriptors[3] can be seen as surveys on feature design, because these features were selected on careful design and comparison to other feature proposals. The currently ongoing SCHEMA project[20] of the European Union intends to provide state of the art reports on content-based media retrieval. At the point in time when this paper is written, deliveries on retrieval concepts, feature extraction and system evaluation are available from the SCHEMA website.

3. THE VISUAL MEDIA

The two types of visual media we are going to consider (image and video) have two major properties that have been examined in VIR research. The first is the colour model used for colour representation and the second is the spatio-temporal resolution of visual media. Colour models have been investigated, for example, by Del Bimbo[2]. Generally, colour models that take human perception into account have been preferred for colour feature extraction. An example is the CIE XYZ space: its unbalanced representation of colours (e.g. more green than red shades) reflects the evolutionary development of the human eye and perception system. For texture and shape analysis, colour models with a luminance channel (originating in TV broadcasting) have been preferred, because, essentially, colour information is irrelevant for this type of analysis. Additionally, a new colour model (HMMD[3]) has been proposed for the MPEG-7 standard. The MPEG-7 authors are arguing that HMMD has properties that make it superior over other colour models. In the author's opinion, since colour values can easily be transformed from any colour model to any other, the selection of colour models is only of minor importance for successful retrieval applications.

Next we will discuss if image and video are similar enough to be handled in one VIR system. The visual media differ significantly in their spatio-temporal resolution. Normally, images have a higher spatial resolution than video. Even though images do usually not contain more information than video frames, due to the different capturing process more scene details are available. The temporal resolution of video is regionally bound and originally derived from TV standards. Images do not have a temporal dimension. Still, a tendency in VIR can be observed to apply features on media objects independently of the availability of a temporal dimension (motion). The authors of the visual part of the MPEG-7 standard stress that their features can be applied reasonably well to both image and video data. They provide structures and models for spatio-temporal localisation and aggregation that allow the application of image features on video content.

We think that in future VIR research the distinction between image and video will become irrelevant. Our argumentation is threefold: Firstly, human vision is a temporal process. The eye scans images and videos by the same saccadic eye movements (to put it simple: close circles in complex areas, larger circles in uniform areas). Therefore, the visual media stream that is sent from the eye to the perception system is always a stream of patterns that has a temporal dimension. Secondly, the result of visual analysis (feature extraction) in VIR is always a number vector of finite length (for technical reasons, etc.). Therefore, image and video are represented by the same type of data. Thirdly, even though some motion features are meaningless for image data, they can at least be used to distinguish the media type by feature vectors. Uniform application of features on media objects is a resource-consuming approach. However, neither

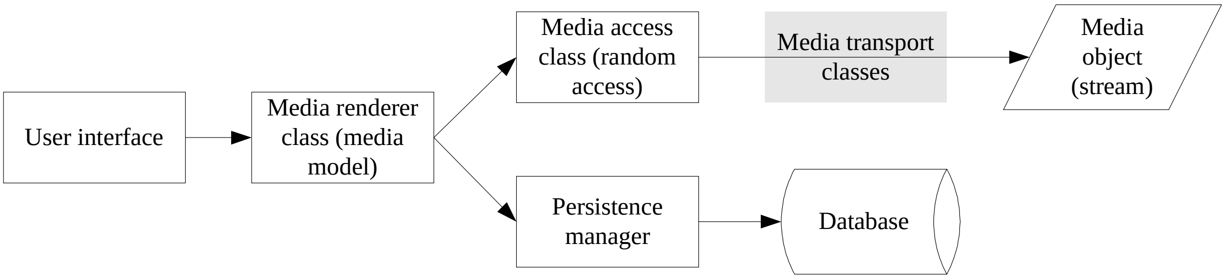

Figure 2: Media encapsulation in VIR.

computation power nor storage is scarce in modern computer systems.

Technically, past VIR prototypes worked either on image data or video data. Mainly, technical shortcomings caused this development. For the future it would be desirable to have VIR prototypes available that support image and video retrieval in a common framework and hide technical media access from VIR-specific tasks (feature extraction, etc.). The author has proposed a VIR framework (called VizIR) that implements this goal[8]. Basically, media access is needed for two functions of VIR systems: feature extraction and media visualisation (e.g. for querying).

VIR video access differs significantly from other media processing applications. Real-time processing is no required. Therefore, video does not have to be considered as a stream but can be accessed like any other pooled data. In the VizIR framework one class is responsible for access of any type of media content. It offers methods for random access of *views*. It is possible to access the view of a media object at any point in time (independent if it is image or video). Additionally, this class is responsible for media content representation and colour space conversion. In a further developed version of this class media objects are abstracted as "visual cubes" (two spatial and one temporal dimension). Transformations (stretching, cutting, etc.) can be applied to manipulate visual cubes.

Media visualisation is (in terms of needed software components) more difficult to perform. The main problem is to visualise the motion in videos in static user interfaces (for querying, result display, etc.). First of all, since user interfaces are normally located on a client while querying components mostly run on a server, media transportation classes are needed that stream the media from server to client. In the VizIR framework, these classes can transparently be attached to the media access class. Media renderer classes are responsible for temporal media visualisation. They make use of the media access interface and construct models of the visualisation that can be used for graphical rendering (e.g. by OpenGL) and be kept persistent in a database. A number of methods have been proposed for video visualisation (e.g. Micons[14]). In VizIR, each method is implemented in a separate media renderer class. Figure 2 summarises the media access components in VizIR.

In conclusion, media-independent availability of visual data in VIR frameworks is a desirable goal. To reach it, making use of software patterns is an important issue (see Section 7). The VizIR framework implements methods for media-independent access. For the future in addition to visual cubes, computing pseudo-saccadic representations of media objects may be worth considering. Completely new features could be designed on the basis of visual pattern streams.

4. FEATURE DESIGN

Since the early days of VIR research, one major focus was on visual feature extraction. The idea of feature transformations is as follows: Since (digital representations of) visual media cannot be easily compared in computer systems (pixel comparison is computational expensive and inadequate to measure *similarity*), there is a need to represent visual content in a form that allows simple but effective (in comparison to human judgement) similarity measurement. In VIR, this is performed by extracting visual media properties as number vectors that can be seen as points in a vector space. If a form of geometry is considered for this space, it is possible to measure dis-similarity as distance. This model is an application of the vector space model of text information retrieval[13].

Since human perception is based on three stimuli: generally perceived (not recognized) stimuli, specifically perceived (recognized) stimuli and pseudo-random (psychological, sociological, etc.) stimuli, two types of features can be distinguished in VIR: *quantitative* (low-level) features and *qualitative* (high-level) features. Unfortunately, only those of the first type can be extracted easily. For the second group semantic understanding would be needed and at the point when this paper is written, software is still far from being able to reason semantically. Therefore, semantic enrichment of low-level features is the mostly adopted course to compute high-level features.

Low-level features are, as pointed out in Section 2, traditionally organised in three groups: (1) colour-related features, (2) texture- and shape-related features and (3) motion-related features. Most colour features (e.g. those in the MPEG-7 standard) extract histograms of pre-defined regions (globally or locally). Only a few approaches exist that make use of colour for other purposes (for example, object segmentation). Texture and shape features can be grouped together, because they make use of the same techniques for feature representation. Both types of features work on the distribution of brightness in visual objects. Texture features aim at detecting statistical edge properties while shape features aim at deriving semantic edge properties (object boundaries). For both types of features it is essential that derived feature representations are invariant against geometric transformations (rotation, scaling, etc.). Motion features include shot detection, camera operation detection and activity detection. Since these features aim at finding features over time, they are mostly built around a core of gradient methods (optical flow, motion trajectories). Usually, low-level feature design results in a cookbook: Building blocks from signal processing (Fourier, Radon transformation, etc.) and other research areas are combined to a new feature. This development has reached a peak in the visual part of the MPEG-7 standard where several cookbooks for low-level features are defined.

One of the most relevant present activities in feature design is semantic enrichment/interpretation of low-level features to narrow the semantic gap. Since as humans we are used to base our similarity judgement on all three groups of stimuli mentioned above, retrieving features just by generally perceived properties is unsatisfactory for us. Generally, three sources of information can be used to enhance features: (1) information on the application domain, (2) information on the user and (3) information on the characteristics of the feature. Additional knowledge can be induced with methods from statistics, artificial intelligence, etc. For example, domain knowledge on football could be used to identify ball and players from shape features (e.g. circularity).

As an example for feature enrichment, in our earlier work we have proposed a semantic feature approach that is based on human perception[9]. Low-level features are used to detect high-level properties that usually play an important role in visual perception. For example, edge and texture features are used to detect symmetries in images. Symmetries are very important for humans. Objects originating from natural processes can easily be distinguished from human-originating objects by their symmetries: Symmetry in nature is never as strict as it is for man-made objects. Probably, it is even possible to distinguish cultures by the symmetries in pictures of their living world. In conclusion, practically, the applicability of semantic enrichment is – at the current point in time – still very limited and for application-independent VIR prototypes no common solution exists.

Another important activity is the ongoing search for 2D segmentation and shape description features. Visual segmentation is the inverse process of rendering. Rendering is a well-posed problem. Therefore, segmentation has to be an ill-posed problem. Nevertheless, the problem is partially solvable, if additional information (on application domain, etc.) is available or if the user helps (for example, by giving a segmentation path). Unfortunately, especially in VIR systems the required additional knowledge (very specific, spatial) is hardly ever present. If we consider, how many different 2D views even a simple object like an apple can have, it becomes unlikely that robust segmentation tools for VIR are possible. However, it will be exciting to see future advances for (narrowly defined) application domains (e.g. salient objects in video).

If we consider the past flood of features, one problem of feature design is obviously answering the question, how many *meaningful* visual features do exist? In other words, which features should be used and which not, because they are outperformed by others? And, on which spatio-temporal regions of media objects should the selected features be applied on? The classic answer to these questions is Multi-Resolution Analysis (MRA). MRA originates in wavelet decomposition. The idea is to make use of a wavelet transformation for computation of wavelet coefficient representations of visual media with decreasing complexity. Either the coefficients themselves or features extracted from the coefficients are used as features (see Figure 3). Unfortunately, it is not clear and could not yet be proven *why* MRA should guarantee that all relevant media parts are properly considered in the feature extraction process.

Our proposal differs from the MRA view: Everything can be a feature, if it fulfils two conditions. Firstly, it has to

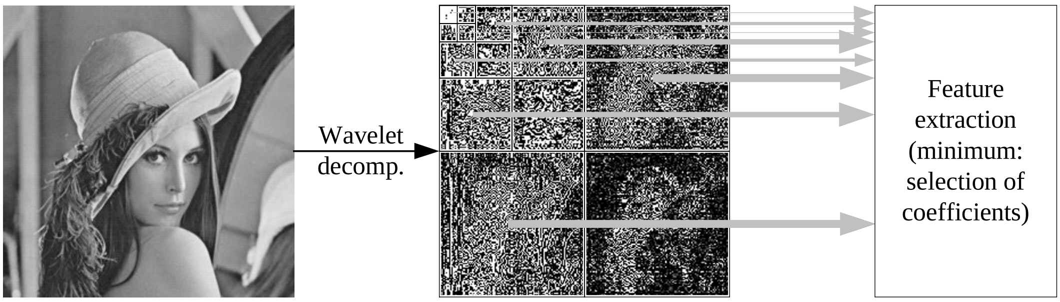

Figure 3: Multi-resolution analysis.

represent a visual property and secondly, it has to be *statistically* independent of existing features. If a feature is statistically independent it is obviously a valuable contribution to a feature set. Independence can be measured by cluster analysis, factor analysis and other methods of statistical data analysis. In previous work we have developed a statistical evaluation procedure and tested the visual MPEG-7 features on these criteria[7, 5] (see also Section 6). Based on this view it is possible to argue for a large number of features to be reasonable. The feature problem is shifted from designing well-performing features to estimating the *relevance* of a feature for a particular querying situation. Essentially, this is up to the user and should be implemented in an iterative retrieval process that makes use of visualisation tools for feature vectors[8].

5. RETRIEVAL PROCESS

Generally, the visual retrieval process aims at finding media objects that are *similar* to given examples. "Similarity" is a weakly defined term and, consequently, difficult to implement in computer systems. Matching by similarity should definitely be less strict than hard pattern matching but still result in comprehensible results. A handful of retrieval processes exists for implementing similarity matching in VIR. Two requirements have to be fulfilled by a model: Similarity matching has to be performed on media objects represented by feature vectors and the user (his feedback) has to be integrated in the retrieval process. Therefore, retrieval is necessarily an *iterative* communication process between man and machine.

Since the actual retrieval process is always based on feature vectors, distinguishing different querying paradigms is irrelevant for the type of retrieval process used. Independently of whether querying by example, sketch, etc. is implemented in the user interface, eventually, the input used for retrieval is always converted to a feature vector (as in text retrieval, where queries are regarded as sets of terms[13]). In consequence we will not refer to different querying paradigms below.

A number of retrieval processes has been introduced to VIR. They are mostly derived from text retrieval concepts. We will consider the four most important models: (1) Distance measurement & indexing, (2) distance measurement and linear merging, (3) distance measurement and non-linear merging and (4) probabilistic retrieval. Except for the last approach, the first step is always distance measurement between the elements of feature space and the given reference point(s).

Distance measurement can be done in two ways: Firstly, a particular type of geometry can be assumed for feature space and metrics can be applied to measure distance. For example, feature space can be assumed to be of Euclidean geometry. Then, the metric axioms hold and any distance measure fulfilling the axioms can be used for distance measurement (e.g. Euclidean distance, city block distance, any Minkowski distance, etc.). Secondly, feature properties (vector elements) can be interpreted as being binary (for example, by fuzzy or probabilistic interpretation). In a binary feature space (populated by binary vectors) predicate-based methods can be used for distance measurement instead of geometric distance measures (e.g. Tversky's well-known Feature Contrast Model[24], Hamming distance, pattern difference).

In recent work we introduced a model that allows for unifying geometric (continuous) and predicate (binary) distance measures[6]. The model allows for using any type of measure on any type of feature data. In experiments on MPEG-7 descriptors we could show that predicate-based measures using the model are often superior over geometric distance measures. The results in the mentioned paper suggest that distance measures should not be designed (derived of feature properties, qualitative arguments) but selected on the basis of quantitative results (e.g. retrieval tests). Generally, the tailor-made distance measure for a feature seldom exists. Optimality depends of the retrieval situation. Therefore, distance measure selection should be automated and derived from given query examples.

Indexing is the art of clever organising data in order to locate them quickly. Since VIR retrieval is based on distance measurement for *all* elements of feature space, indexing as an acceleration technique is irrelevant for querying. But indexing can be used as a querying method itself. In high-dimensional index structures those regions can be selected as positive retrieval results that lie in proximity to the given examples. Unfortunately, hardly any indexing methods do exist that could deal with multiple distance measures and variable (in terms of query examples) data organisation. Therefore, the applicability of indexing methods for VIR is relatively limited.

Linear and non-linear merging approaches are addressing the problem of how to use multiple features (and distance measures) in one query and to retrieve single result set. Linear merging solves the problem by weighting the distance values and summing them up for each media object. Next to weights, transformations are used as well. The resulting value is used to rank media objects and select the first ones as similar. Two problems are connected to linear merging: the weights and the size of the result set have to be provided by the user and some features cannot be combined linearly. Non-linear merging tries to overcome these problems. Often, neural network techniques are used to combine individual distance values to a rank. For example, a multi-layer feed-forward net can be trained on basis of ground truth information. Unfortunately, non-linear methods are – as any other retrieval method – not able to satisfy all user needs and are hardly configurable because of their inflexible architecture.

Using probabilistic approaches (for example, the Binary Independence Model[13]) for retrieval results in two major problems. Firstly, since most models where developed for text retrieval they require binary input that is seldom available in VIR. Again the same methods as for predicate-based distance measurement can be used to convert continuous values to predicates but every additional interpretation step reduces the quality of the results. Secondly, probabilistic models judge general relevance (similarity) on basis of elementary (feature-wise) relevance information. This relevance information has to be provided in form of examples. Already difficult for text retrieval this is nearly impossible for visual data, because the number of possible features and feature values (representing all types of visual cues) is nearly indefinite. Therefore, if probabilistic model are used, then mostly in elementary form (e.g. simple applications of Bayes' theorem).

One major advance in VIR in recent years was achieved in iterative refinement by relevance feedback. Clearly, retrieval should be centered around the user but the question arises of how to apply his feedback in the retrieval process. Here, kernel-based learning techniques[17] mark a significant advance. Using results of previous queries that are enriched by elementary user feedback ("highly relevant", "irrelevant", etc.) as reference points and training a kernel function to segment feature space optimally improves results dramatically. After all, finding a dichotomy of relevant/irrelevant media objects is all that is required of a VIR system. Often used kernel-based learning methods include support vector machines and kernel principal component analysis. The main problem of applying kernel-based learning to VIR is finding a kernel functions that neither over-fits (too complex, too high dimensionality) nor under-fits (too simple, bad segmentation) the retrieval problem.

Unfortunately, even the most sophisticated retrieval and refinement algorithms are still not able to satisfy the user's desire for similarity-based retrieval sufficiently. Therefore, we have designed a retrieval process (called *visual mining*, VM) that is user-centered from the first to the last querying iteration and makes use of 3D perception. Figure 4 shows the retrieval process schematically. Media objects are visualised on the image plane while on the floor dimensions their relative location (distance) is visualised for two features. The features selected for the floor dimensions can be changed at any time implying changes in the organisation of the media objects. This form of visualisation allows the user to visually perceive the retrieval process. Queries are defined by labelling media objects as positive or negative examples. Implicitly, the labelling defines hyper-clusters. The query engine tries to fill the defined clusters with similar objects. For this purpose it makes use of distance functions and data segmentation methods.

Visual mining aims at really putting "the human in the loop"[18]. In Figure 4 image and video objects (represented as Micons[14]) are used in the same query. In a typical querying situation multiple instances of the shown panel are used. For example, one for query definition, one that shows the last result set, one that gives a general overview over feature space, etc. The VM process and the user interfaces are explained in more detail in recent publications[11, 10].

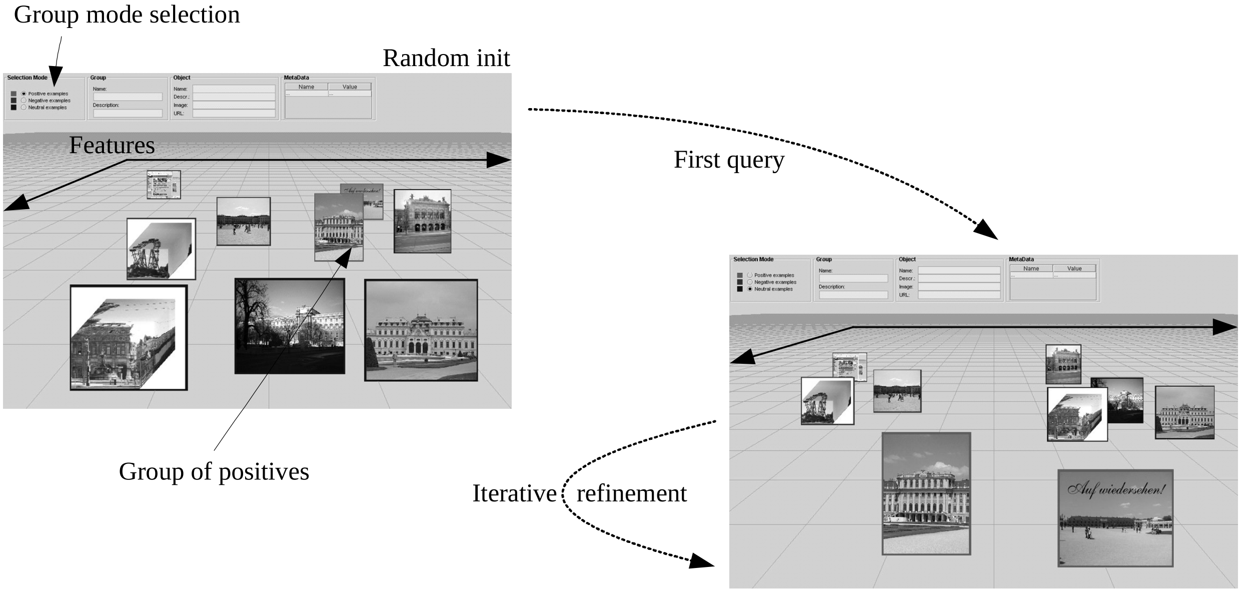

Figure 4: Iterative group querying process.

In conclusion of Section 4 and 5, a great variety of feature design and VIR retrieval methods exists that all have their advantages and disadvantages. To be useful for practical application it is necessary to be able to judge the specific qualities of querying prototypes. In the next section, the methods mostly used for VIR evaluation are shortly sketched and new methods that could supplement existing ones are proposed.

6. EVALUATION

Evaluation of VIR systems is needed for various purposes: It has to be possible to judge the quality of new feature extraction methods in relation to existing ones, to compare the quality of novel querying paradigms, to judge the usability of user interfaces for retrieval, etc. The most interesting problem is measuring the quality of similarity measurement compared to human visual similarity perception. For this purpose, the recall and precision quality indicators of text information retrieval evaluation were adopted[13]. Recall and precision are usually defined as follows:

$$recall = \frac{|retrieved \cap relevant|}{|relevant\ objects|}, precision = \frac{|retrieved \cap relevant|}{|retrieved\ objects|} \qquad (1)$$

In case of VIR, objects are media objects represented by feature vectors. Recall and precision are inter-dependent. It is easily possible to optimise one indicator, if the other is not considered. Meaningful results can only be derived if both indicators are considered. In addition to recall and precision other measures exist (for example, ANMRR, used for evaluation of visual MPEG-7 descriptors[15]).

VIR evaluation based on recall and precision is a four-step process (see Figure 5): (1) Definition of a media set. The media set should be appropriate for the evaluation goal and contain a reasonably large number of items. Often, collections of thousand and more media objects are used. (2) Derivation of ground truth information. The ground truth says, which objects in a media set are similar (and sometimes, *how* similar they are). Ideally, it should be invariant against cultural, sociological and other human-related influence factors. In practice, deriving such a ground truth is impossible. Usually, groups of more than average similarity are defined by a few test users. (3) Execution of test queries. This step requires automatic selection of query examples and a sufficiently large number of test queries. For guaranteeing statistical correctness, the number of test queries should be hundred or larger. (4) Computation of retrieval indicators. Recall and precision can, for example, be averaged over all test queries and visualised in a recall-precision-graph. This evaluation procedure has several shortcomings: Firstly, it is subjective and culture-dependent (media

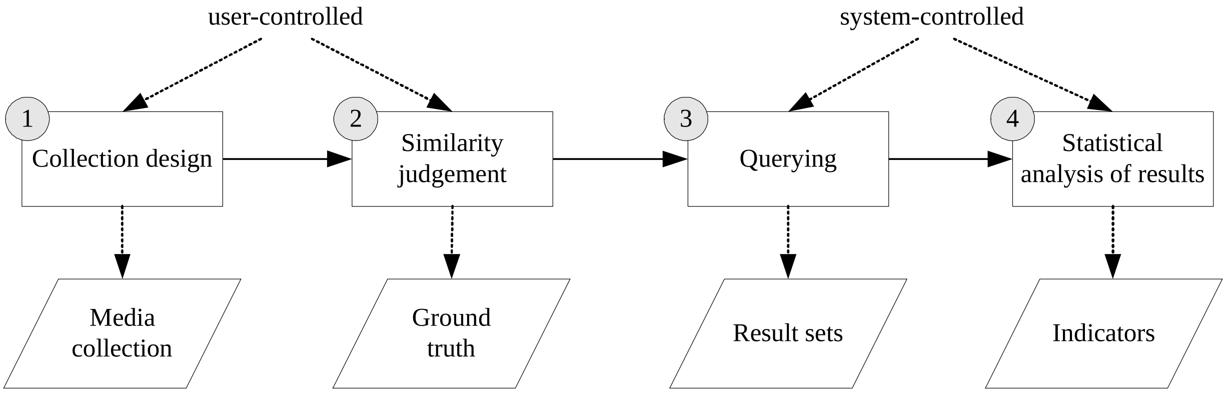

Figure 5: VIR evaluation process.

collection, ground truth). Secondly, it cannot be used to evaluate interactive retrieval processes. Thirdly, it is a heavy-weight process that adds a lot of influence factors that may bias the evaluation results. For example, this may be the case if a new feature should be evaluated.

Present evaluation activities include gathering free media objects in public collections (e.g. the Benchathlon project[1]) and events for comparative system evaluation. One example for the second is the annual TREC video retrieval competition[21]. VIR groups can attend in a number of competitions (e.g. shot segmentation) and see how good their methods are in comparison to other approaches. Additionally, a new (very large) set of video clips is created each year that can be used for other purposes as well. This is especially positive since most freely available visual media collections are image collections.

In our recent work we have proposed an evaluation procedure for features that is based on statistical data analysis and the visual MPEG-7 features[5, 7]. The procedure makes use of factor analysis and cluster analysis techniques. In contrast to the standard procedure it does not suffer from the three mentioned disadvantages. Essentially, feature vectors are calculated for arbitrary media collections and compared to the MPEG-7 feature vectors by statistical methods. The results can be used to judge the feature type (colour, texture, etc.), redundancies with existing approaches, etc. It is intended to be used as a supplement to recall- and precision-based evaluation.

7. SYSTEM DESIGN

Good, professional system design is not a VIR-specific issue; it is desired for any type of information system. What makes system design especially important in VIR is the fact that acceptance of VIR methods is strongly bound to their appearance. Since VIR systems actually fail to fulfil the promise of human-like similarity retrieval, it is even more important that they are at least fast and easy to use tools for visual media mining (pre-selection of likely hits). Below, we point out the design of classic systems, currently ongoing design activities and our ideas for ideal VIR system design.

Past VIR prototypes were mostly monolithic systems that ran on server side and were limited to one type of media. Most VIR systems implemented image retrieval: a few features (colour histogram, texture moments, etc.), query by example and retrieval by linear merging. Most of them were general-purpose, some application-specific (e.g. for trademark retrieval). Video retrieval systems were mostly intended for specific applications (e.g. news analysis) and often concentrated on the user interface aspect (visualisation of temporal media in static user interfaces). Well-known VIR prototypes include QBIC, Virage, RetrievalWare, Photobook, VisualSEEk, MARS, OVID and CueVideo. Surveys exist that evaluate these and other prototypes and compare them by their advantages and disadvantages[16, 27].

IBM's Query by Image Content system[12] (QBIC) may stand as a representative for these prototypes. QBIC is a classic system that introduced many of the concepts that are implemented today in a wide range of VIR prototypes. QBIC is based on the C++ programming language and organised in components. The architecture is extendible: new features and query engines can be defined and added. Querying components are separated from the user interface and communicated over HTTP. Image data is encapsulated in a data class that is also responsible for converting various image formats to

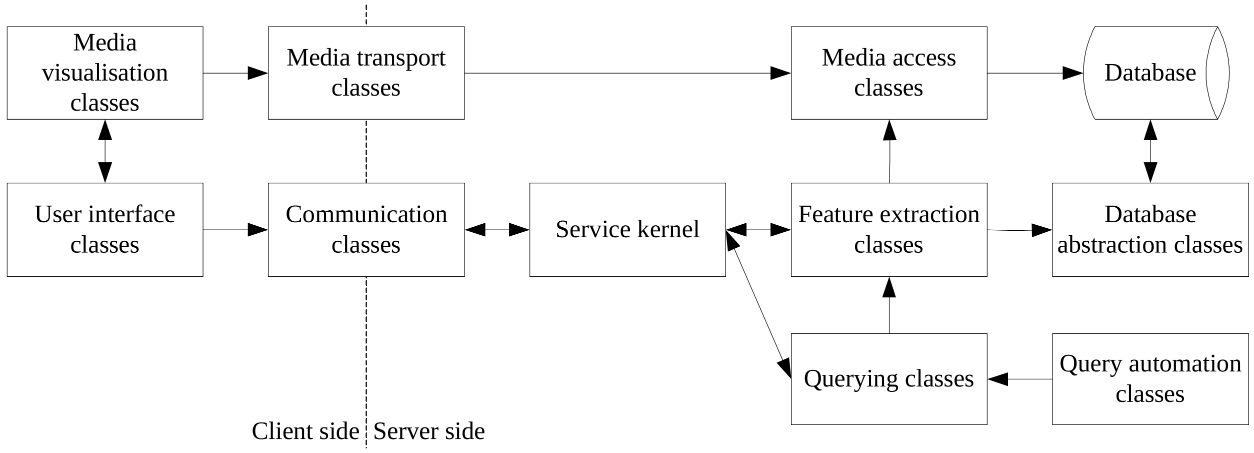

Figure 6: Ideal VIR system design. Arrows show "make use" dependencies.

raw RGB pixel maps. Those source code elements needed for the extension mechanism are shipped with the binary distributions for various operating systems. QBIC contains a number of state-of-the-art feature classes and used linear merging for retrieval. Additionally, it is based on a simple file database for feature storage.

At present, these concepts are imitated in a number of prototypes. For example, the GNU Image Finding Tool[25] (GIFT) makes use of the Multimedia Retrieval Markup Language[26] (MRML, based on XML) for loose coupling of server and client components. GIFT is open source and based on other GNU components that allow using a large number of data formats for image querying. Since the communication language for server and client components is standardised, different user interfaces can be used to access the query engine.

The MPEG-7 experimentation model[23] (XM) goes one further step ahead, as it allows querying in image and video collections. Like for QBIC and GIFT, the XM classes are split in server components (for querying) and client components. It allows extension with new descriptors and is available as open source. Unfortunately, the practical use of the XM is limited, because only a very small number of video formats are supported and hardly any documentation exists for architecture and application programming interfaces. Still, the XM is used as basis for a number of VIR projects. For example, the SCHEMA project of the European Union[20] develops new VIR solutions on basis of the MPEG-7 XM. Other projects (e.g. of the DELOS Network of Excellence of the European Union[4]) are following different, individual approaches.

In recent publications we have proposed an "ideal" architecture for VIR systems. This architecture is currently under development in the VizIR project[11]. One major goal of the VizIR project is providing a framework of VIR tools that are media-independent. Another is encapsulating visual media in a way that most common image and video formats are supported and that media content can be accessed with exactly the same methods. VizIR is an open source project that is based on the Java programming language. It implements all of the proposals for feature design, retrieval and evaluation made in this paper.

Figure 6 shows the VizIR system design. Components are split into typical client components (user interfaces) and server components. Client components are the user interface presented in Section 5 and the classes for visual media representation presented in Section 3. On the server side a service kernel is responsible for dispatching server calls (e.g. query execution, media management). This service kernel can, for example, be implemented as a web service using SOAP, WSDL and UDDI. It organises the classes for querying and feature extraction that are derived from general interfaces. Therefore, it is easily possible to extend the VizIR framework with new features and querying paradigms. Database storage and additional functionalities for query acceleration (feature vector indexing, querying heuristics, etc.) are encapsulated in an object-oriented persistence manager that hides the database (for feature storage, etc.) from the VIR-specific classes. The same purpose is fulfilled by the media access classes for the media objects. Query automation classes are used for evaluation purposes.

Communication between server and client side is performed by communication classes that make use of XML messaging and are fully compatible with the service kernel. For media transport individual classes are implemented that fulfil their job in separate threads in the background. It is important to notice that all VizIR framework components are designed to be applicable independently of the type of media used and of the location from where they are used. It is possible to build arbitrary VIR applications by using existing building blocks. New ones can be added easily. In order to guarantee that every component can communicate with any other, event-based messaging is used and implemented following established design patterns (e.g. SUN's delegation event model). Generally, design patterns are used wherever possible (e.g. factories for media access).

8. CONCLUSIONS & OUTLOOK

This paper summarises selected advances in visual information retrieval. We try to sketch important advances in visual media representation, feature extraction, retrieval (including query definition, similarity measurement and query refinement). Additionally, we propose problem areas and possible solutions for future visual information retrieval research. The selection is subjective: it represents the author's point of view on image and video retrieval.

The major problem of visual information retrieval is its failure to imitate human visual perception and human similarity judgement properly. The goal is to automatically find visual media in, usually very large, collections by imitating human visual similarity perception. Clearly, since computers are still unable to do visual reasoning and recognise the real world objects behind two-dimensional views, they are condemned to fail. What they can do is to extract visual features on a low syntactical level and to measure dis-similarity as distance. Even though this service can be of great value (e.g. as a pre-selection step when mining large media collections), the unsatisfactory results are a major reason why content-based retrieval techniques are still hardly used in digital library systems and other applications.

In consequence, the key question is: does visual information retrieval have a perspective for practical application? To the author's belief, this question can be answered by "yes" if research and implementation focus are laid on issues different from the currently most investigated. Visual information retrieval is a mining tool that should be centered around the user and have its major strength in the user interface components used for media and query visualisation. Systems have to be designed in an easy to use way and it has to be made clear that visual information retrieval systems are not intended to replace but to supplement human beings and their visual perception system.

ACKNOWLEDGEMENTS

The author would like to thank Christian Breiteneder for his valuable comments and suggestions for improvement. This work is part of the VizIR research project[11] that is supported by the Austrian Scientific Research Fund (FWF) under grant no. P16111-N05.

REFERENCES

1. Benchathlon network website, http://www.benchathlon.net/ (last visited 2003-10-25).

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

3. 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.

4. DELOS EU Network of Excellence website, http://delos-noe.iei.pi.cnr.it/ (last visited: 2003-10-25).

5. H. Eidenberger, "A new method for visual descriptor evaluation", *Proceedings SPIE Electronic Imaging Symposium*, SPIE, San Jose, 2004 (to appear).

6. H. Eidenberger, "Distance Measures for MPEG-7-based Retrieval", *Proceedings ACM Multimedia Information Retrieval Workshop*, ACM Multimedia Conference Proceedings, Berkeley, 2003 (to appear).

7. H. Eidenberger, "How good are the visual MPEG-7 features?", *Proceedings SPIE Visual Communications and Image Processing Conference*, vol. 5150, 476-488, SPIE, Lugano, 2003.

8. H. Eidenberger, "Media Handling for Visual Information Retrieval in VizIR", *Proceedings SPIE Visual Communications and Image Processing Conference*, vol. 5150, 1078-1088, SPIE, Lugano, 2003.

9. H. Eidenberger, C. Breiteneder, "Semantic Feature Layers in Content-based Image Retrieval: Implementation of Human World Features", *Proceedings IEEE International Conference on Control, Automation, Robotic and Vision*, Singapore, 2002 (published on CD, available from: http://www.ims.tuwien.ac.at/~hme/papers/icarcv2002.pdf, last visited: 2003-10-25).

10. H. Eidenberger, C. Breiteneder, "Visual similarity measurement with the Feature Contrast Model", *Proceedings SPIE Storage and Retrieval for Media Databases Conference*, vol. 5021, 64-76, SPIE, Santa Clara, 2003.

11. H. Eidenberger, C. Breiteneder, "VizIR – A Framework for Visual Information Retrieval", *Journal of Visual Languages and Computing*, **14**, 443-469, 2003.

12. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, "Query by Image and Video Content: The QBIC System", *IEEE Computer*, **28/9**, 23-32, 1995.

13. N. Fuhr, "Information Retrieval Methods for Multimedia Objects", *State-of-the-Art in Content-Based Image and Video Retrieval*, R.C. Veltkamp, H. Burkhardt, H.P. Kriegel, 191-212, Kluwer, Boston, 2001.

14. B. Furht, S.W. Smoliar, H. Zhang, *Video and Image Processing in Multimedia Systems*, Kluwer, Boston, 1996.

15. 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.

16. O. Marques, B. Furht, *Content-Based Image and Video Retrieval*, Kluwer, Boston, 2002.

17. K.R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, "An Introduction to Kernel-based Learning Algorithms", *IEEE Transactions on Neural Networks*, 12/2, 181-202, 2001.

18. Y. Rui, T.S. Huang, S.F. Chang, "Image Retrieval: Past, Present, And Future", *Proceedings International Symposium on Multimedia Information Processing*, 1997.

19. S. Santini, R. Jain, "Similarity Matching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21/9**, 871-883, 1999.

20. SCHEMA EU project website, Delivery on visual information retrieval techniques, available from http://www.iti.gr/SCHEMA/preview.html?file_id=67/ (last visited 2003-10-25).

21. A.F. Smeaton, P. Over, "The TREC-2002 video track report", *NIST Special Publication*, SP 500-251, 2003 (available from: http://trec.nist.gov/pubs/trec11/papers/ VIDEO.OVER.pdf, last visited: 2003-10-25).

22. 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.

23. TU Munich, MPEG-7 experimentation model website, http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/ mmdb/e_mpeg7.html (last visited: 2003-10-25).

24. A. Tversky, "Features of Similarity", *Psychological Review*, 84/4, 327-352, 1977.

25. University of Geneva, GNU Image Finding Tool website, http://www.gnu.org/software/gift/ (last visited: 2003-10-25).

26. University of Geneva, Multimedia Retrieval Markup Language website, http://www.mrml.net/ (last visited: 2003-10-25).

27. R. Veltkamp, M. Tanase, D. Sent, "Features in Content-based Image Retrieval Systems", *State-of-the-Art in Content-Based Image and Video Retrieval*, R.C. Veltkamp, H. Burkhardt, H.P. Kriegel, 97-124, Kluwer, Boston, 2001.