Preface

Images and video play a crucial role in Visual Information Systems and Multimedia. There is an extraordinary large number of applications of such systems in entertainment, business, art, engineering, and science. Such applications often involve a client-server architecture, with large file and compute servers. Searching for images and video in large collections is becoming an important operation. Because of the size of such databases, efficiency is crucial.

We strongly believe that image and video retrieval need an integrated approach from fields such as image processing, shape processing, perception, data base indexing, visualization, querying, etc. On the other hand, most ongoing projects only deal with one or two of these aspects. A research emphasis is needed on incorporating multiple models for shape, color, texture, geometry, and syntax, so that the user does not have to specify low-level model parameters and combinations. This should lead to strategies of efficient indexing of visual information, in addition to techniques for combining visual with more traditional database information. Realistic evaluation criteria are needed, including test databases of realistic size in domains of interest, measures of similarity that allow variations in perceptual, semantic, and other criteria, and measures of accuracy and efficiency in assisting the user.

The purpose of this first Dagstuhl Seminar "Content-Based Image and Video Retrieval" was to bring together people from the various fields in order to promote information exchange and interaction among researchers who are interested in various aspects of accessing the content of image and video data, including topics such as:

- Indexing schemes
- Matching algorithms
- Visual data modeling
- Retrieval system architectures
- Image and video databases
- Feature recognition
- Video segmentation
- Picture representation
- Query processing
- Perception issues
- Video and image compression
- Visualizing pictorial information
- Searching the web
- Delivery of visual information
- Benchmarking
- Application areas of image and video retrieval

For this seminar, we have invited internationally known as well as young researchers from various disciplines with a common interest in content-based image and video

retrieval. We have been together with a group of 28 researchers for a week, away from the rest of the world, and certainly good interaction and exchange of ideas took place during the sessions as well as in the very "gemütliche" wine cellar, enjoying the cheese platter.

There was a total of 24 presentations, two demonstration sessions, and two discussion sessions. The abstracts of the presentations are presented here, in the order in which they have been written by hand in the 'book of abstracts'. One discussion session was about the challenges and problems to be solved, the other session was about the particular problem of how to assess the quality of retrieval systems and algorithms for subtasks, see the report at the end of this booklet.

Hans Burkhardt, Freiburg Universität
Hans-Peter Kriegel, Universität München
Remco Veltkamp, Utrecht University

Toward Interactive Time Similarity Search
Roger Weber, ETH Zürich, Switzerland

A pressing problem studying between a large image database and its usage is how to effectively and efficiently satisfy a user's information needs. A common approach to search an image database is content-based retrieval. The problem is that this kind of method is very costly: (1) we have to extract features, (2) we have to query indexes, and (3) we have to deliver the images. We have addressed all the problems but for the sake of time we only presented solutions to the second problem. Our main approaches are:

- Parallelization (several disks / processors, batching, a cluster of workstations).
- Approximate NN - search (trade response time for result quality).
- Refining / adapting a query such that only the cheap indexes are considered.

Our experiments with our database (>100.000 images) indicate that, with the above measures, index lookup is no longer the bottleneck of the system.

Adaptable Similarity Search in Image Databases
Thomas Seidl, Hans-Peter Kriegel, Univ. Munich, Germany

Similarity search is a highly application-dependent and even subjective task. Similarity models are derived to be adaptable to application specific requirements or individual user preferences. In our work we focus on two aspects of adaptable similarity search.

- Adaptable Similarity Models. Examples include pixels based shape similarity, 2D and 3D-shape histogram, and shape oriented similarity models. These models are applied to mathematics, biomolecular and medical image databases.
- Efficient Similarity Query Processing. Similarity models based on quadratic form result in ellipsoid queries in high-dimensional data spaces. We present an algorithm to efficiently process ellipsoid queries on index structures, and to improve performance by introducing several approximation techniques while guaranteeing no false dismissals for similarity range queries and k-nearest neighbor queries.

Color Vision and Image Search Systems
Arnold Smeulders, Theo Gevers, ISIS group
University of Amsterdam

Before considering an image search it is important to make a distinction by

- The domain (is it narrow or broad).
- The sensing environment (is it 2D of 2D scene, 3D of 3D scene, or 2D of 3D scene; is it loose in the scene or occluded and cluttered).
- The type of query (1 to find the best, *n* to find the best, interactive definition of the query).

- The type of variances / invariances suitable for the query.

This information can only be provided by the user.

We have provided a set of color invariant features suitable for a variety of different circumstances.

For shape we propose

- Distinction between geometric shape, requiring successful segmentation.
- Direct shape comparison, without the use of shape features.
- Shape as a differential geometric property of color scale space.

New Descriptors for Image and Video Indexing
Patric Gros, Vista Group, IRISA CSNRS, France

In this talk, I gave a survey of research works driven in the Movi group at INRIA in Grenoble where I was not too long ago and in the VISTA group where I am currently.

The Movi group is interested in the problem of finding invariant or quasi-invariant descriptors to match images and solve the object recognition problem. This approach is based on a point extractor, version of Harris detector with varying parameters to have a multi-scale approach, on the computation of the local jet (see Florak et. al.) and on the mix of the obtained derivatives to get invariants. An extension to the case of color was presented, as well as another extension to the case of video, where image understanding can be used to improve the invariant robustness.

The VISTA group is interested in the specific problems that arise using image segments. The group has developed a method to compare a parametric model of the global motion between two images and of the image data corresponding to that motion. This motion field is then used to detect cuts, dissolves and wipes, to detect and track moving objects, to hack any region given by the user. From any motion field, it is also possible to compute global motion descriptors that can be used to retrieve sequences according to their global activity level.

Video Retrieval by Content
Alberto Del Bimbo Univ. of Florence Italy

Video retrieval systems are highly dependent in the type of video. They go through three distinct steps: segmentation (shots/episodes/scenes); annotation (either normal or based on speech/text); retrieval (including querying browsing and visualization). Access to video must be perceptual at some semantic level in order to match user's expectations when he sees a query.
We now discuss those research topics that we have concluded useful:

- *Retrieval based on semantically meaningful categories (of advertising).* Perceptual features like presence of cuts / dissolves, salient colors straight/horizontal/vertical lines, resemblance of colors are useful to group films into 4 semantic categories (playful, critical, utopic, practical). We present examples of a retrieval system which allows retrieving advertising quickly features of one or more categories and retrieve advertising films that are similar to a query film according to its semiotic content.
- *Retrieval based on text (of news).* In news, content is particularly represented by words. Images have in fact an auxiliary function with respect to text. We have developed a system which segments the film into shots, detecting the anchorman shot and extracts from this shot text (in textual stripes) or text from speech (through speech recognition). The anchorman shot is recognized without an anchorman model but simply based on semiotics of absence of motion.
- *Retrieval based on color flows (of advertising).* Content of video search represented as a set of flows of color regions. These color flows are the more responsible of our perception of the film content, together with words. In advertising where words are of little significance, they are a very meaningful representation of context at the perceptual level. To extract flows we cluster colors in each frame, identifying color regions, and track color regions in subsequent frames. Flows are encoded through a multidimensional Haar wavelet transform. Retrieval of films using similar flows is implemented by matching wavelet coefficients, flow volume similarity, color similarity, and semiotics similarity.

Supporting Image Retrieval by Database Driven Interfaces to 3D Visualization

Erich J. Neuhold, GMD-IPSI, Germany

Supporting image retrieval between naive users and information systems requires careful design of the visualization and interaction components that in many cases also depends on the context in which the search is executed. We have selected to illustrate the possible advantages of a coupling between interactive 3D visualization systems and image retrieval systems based on database management systems. We have used as a characterizing application scenario an interactive 3D virtual gallery. The talk offered an analysis of the requirements of components and architecture of a generic database driven 3D visualization system based on RDBMS and VRML technology. Special emphasis was given to the objective of buying and selling objects of art through a virtual gallery metaphor, where the retrieval results will be displayed in a building, an environment with floors and rooms. Searches are currently based on textual indexing of the objects of art, the visualization takes place in several levels of detail, and decision to buy a piece of art can then level to a fully electronic of traditional process.

Currently, efforts are underway to extend the search functionality to content based retrieval. Many techniques exist but it is already clear that a careful selection needs to be

made between them in order to satisfy the application-specific requirements (art market) the sellers and buyers reflect.

Content Based Image Retrieval Using Statistical Functions
Andre Everts, GMD-IPSI, Germany

Current research and development in image retrieval has yielded an extensive collection of feature extraction algorithms. Just to mention a few, textural analysis, shape comparison, color similarity etc. have shown to be useful matching tools in certain domains. This talk describes a formal framework to provide means for effective, interactive retrieval sessions. The key idea is to model the characteristics of image retrieval algorithms. Each algorithm is described as a rule, based on feature values, which reflects its average performance in a manually classified training set of randomly selected images. First the rules where developed by a brute-force algorithm which shows a good behavior. Later we changed to statistical rules to cluster the content of an image. An evaluation of the rules depending on a test set of manually classified images was also given in the talk.

Local Features for Recognition and Registration
Luc Van Gool, Univ. Leuven

Pictures and construct of model based and appearance-based approaches to object recognition seem quite opposite. This may indicate the combination of both schools holds good promise. In my view the way by Schmidt and Mohr is a good step in that direction. In our talk we try to further extend the specific advantages of their approach. One aspect is the extension of the level of invariance by which their interest regions are discussed. From geometric to geometric and photometric and from geometric rotation in the image plane to full plane affine invariance. Key to such extension is the construction of invariant neighborhoods with self-adaptive shapes: under changing viewpoint they always should be selected independently from other views, to cover the same physical part of the scene. In the meantime several of such projections for their extraction have been derived. Results for database retrieval are shown.

Distribution Based Image Similarity Measures
Jan Puzicha, UC Berkeley

The aim of this talk is two-fold: First I present a novel benchmark methodology to compare distribution based similarity measures based on color and texture. An extensive overview of current similarity measures is given and a categorization is achieved.

Ground truth is defined by a novel sampling scheme where a single image defines the source of a class. Results are given for classification, retrieval and unsupervised segmentation demonstrating that a careful selection of a measure substantially increases performance. I conclude by stating that there is no overall best measure but rather different tools for different tasks.
In the second part of the talk, I introduce a novel retrieval model referred to as probabilistic image explanation. Similarity is defined by the log-posterior probability of a database model. It is shown how fast retrieval can be achieved by a novel sequential estimation technique. The novel measure outperforms other similarity scores and gives excellent results in practice. Its open system architecture is highlighted by an integration of a color and texture database.

Statistical Approach to Intensity Based Object Recognition
Heinrich Niemann, Univ. Erlangen-Nürnberg

Three approaches to obtain a statistic model of an object are presented briefly: the probability of an intensity (or some function of it) conditioned by position in the image; the probability of a position conditioned by intensity; the probability of segmentation results. All models depend on the object class and object pose in order to allow recognition and localization.
The case "probability of position by conditional intensity" is considered in more detail. The main assumptions are statistical independence of positions, mixture of normal distributions and coarse quantization of intensities. Unknown parameters are estimated by the EM algorithm with is initialized by vector quantization. Localization is done by global optimization of the likelihood functions. Classification is by Bayes rule.
Classification and localization experiments were performed with four objects one of them of irregular shape (a cactus). Only 512 pixels out of the whole image matrix are sufficient to reliable locate and classify (approx. 94%) the objects.

Image Retrieval by Orientation Field
Josef Bigun, Halmstad University, Sweden

Retrieval of images in databases enabling further study with respect to their contents is at the focus of attention. But first the idea of orientation fields in connection with human visual system is presented. Gabor filters that approximate the linear behavior of the simple cells are presented. However, Gabor filtering is computationally expensive. Linear symmetries, an alternative way of getting orientation fields, are presented.
The main difficulties of image retrieval consist in processing the images rapidly. We propose orientation radiograms to be used as image signatures for shape based queries. These are projections of a set of orientation decomposed images to axes whose directions change synchronously with the orientation bands at hand. We present the results concerning 100 ornaments in printed old books, which is at the focus of attention of book historians. Also a comparative study is presented comparing classical moment invariants.

Facial and Motion Analysis for Image and Video Retrieval
Massimo Tistarelli, DIST, Univ. of Genova, Italy

Among the many visual cues, which may be of use for image and video retrieval, two play a very special role: motion and face images.

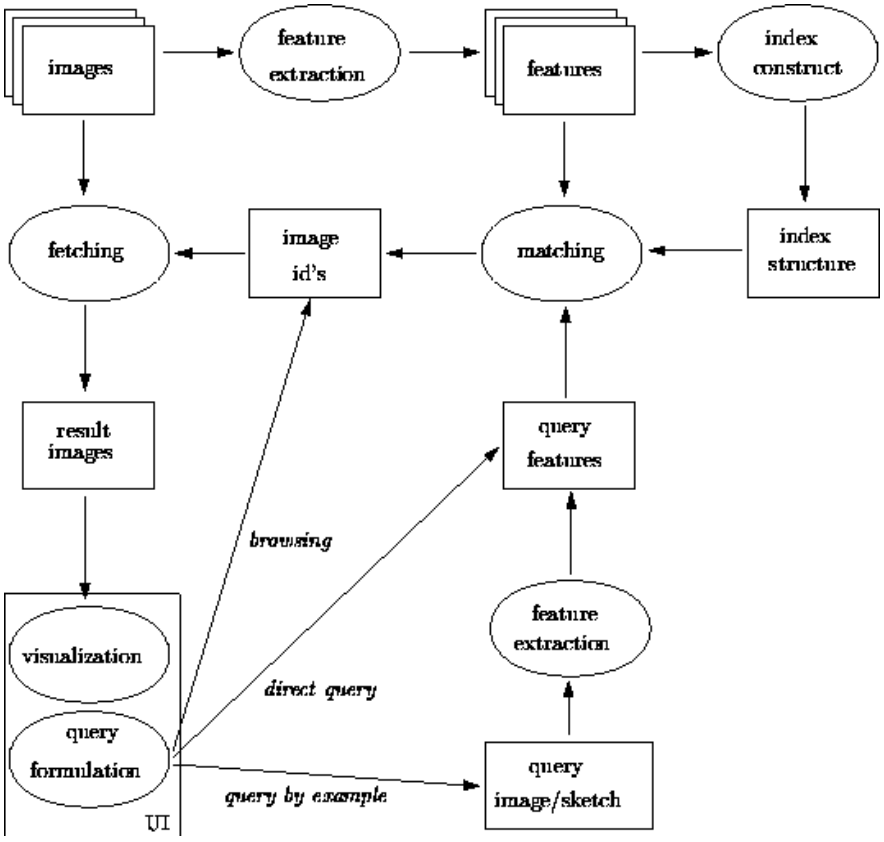
Motion itself conveys a lot of information about the content in image sequences and sometimes, this is the most distinguishing features (for example in short movies). A very popular representation for image motion is optical flow, which is an approximation of the 2D projection of 3D motion on the image plane, for every image pixel. It has been enormously believed that optical flow computation is difficult or not practical because of the ambiguities rising from the so-called "aperture problem". It is proven that the opposite problem holds: there is an abundance of constraint expressions which can be used to compute image velocity, but, in order to gather the correct result, you have to carefully select the most appropriate constraints. It turns out that, starting from a conventional gradient-based approach, that local gray level structure in the image determines which constraint equations should be applied. Optical flow per se conveys too much data to index video data, but a concise description of the events can be extracted either by defining a few elementary motion (or flow patterns) and fitting a parametric description to the flow field, or by decomposing the flow field into an elementary set of basis vector fields.

Facial images share many common properties that allow the implementation of efficient retrieval systems. Classical, most famous approaches based on the SVD decomposition of an image ensemble are not practical for scenarios where data is recorded in different formats and updated dynamically. On the other hand these methods rarely rely on features which are specific for face images. In general, the problem is to provide a classification among different sets, each containing several instances (pictures) of the same subject. The dimensionality of the problem makes it very difficult to use classical classification methods such as Bayesional classifiers or Support Vector Machines. A possible alternative is to define the separation between the subject to be retrieved and "the rest of the world". Even though this option may seem more complex than the general recognition, it turns out that in this way the parameterization is limited to the face image used for the query. Of course, the parameterization of the face must be carefully chosen to limit a presence of outliners in the representation. An example of this technique implying a log-polar representation of facial features and correlation is shown. With this methodology a classification error of about 2% has been achieved, which drops down to 0.5% when integrated with a similar technique based on space-variant Gabor representation and SVM classifier.

Surveying the World of SMURF

Remco Veltkamp, Daniëlle Sent,
Utrecht University, The Netherlands

In this presentation we survey the functionality and implementation of temporary content-based image retrieval systems. Of course these have been made other overviews, but here we concentrate on technical aspects. We do this in the basis of the following similarity-based multimedia retrieval framework (SMURF).



In particular we investigate which kind of color, texture, shape and spatial layout features are used, how the matching function is done, whether indexing structures are used, how the query formulation takes place, and how the results are visualized. Apart from tabulating the capabilities of existing systems, the main conclusion is that we need a means to evaluate and compare performance of systems and propose to set up an evaluation experimentation framework by the community, and hold a contest or competition.

Using Invariants for Content Based Image Retrieval Fundamentals and Applications, Part I Haus Burkhardt, Institut für Informatik, Univ. Freiburg

The contribution describes the use of invariants for image retrieval purposes. We consider images to be equivalent if they can be generated out of each other by Euclidean motion (translation and rotation) of the whole image as well Euclidean motion of individual object in the scene. The invariants are generated on the basis of their integrals over the transformation group in the equivalence class. By using kernel functions for the integration of local support the extracted features are well tolerant to variations from architectural objects well as topological deformations. The theoretical fundamentals are covered and several examples are given: image queries, visual implementation of textile banner and flame classification as well as the extension to 3D data (medical tomographical volume data, pollen classification).

Image Content Analysis and Description Xenophontas Zabulis, ICS FORTH, Greece

We aim towards a qualitative description of image content out of quantitative estimations of image feature values. In order to do so we study perceptual and biological models of human image understanding.
We use the shape, texture, and color image features as well as their spatial layout as a first primitive and quantitative image description. Through this study we try to organize features into perceptual groups, which is shown to be a qualitative image feature. Also the study of human visual attractors is employed in order to detect attention attracting image regions, which are regarded as qualitative image features as well. In order to obtain an image content representation both abstract and detailed a multi featured scale space is considered to be used. Multiple feature integration is considered to be implemented through the use and evaluation of several voting systems.

Color and Spatial Feature Extraction Jean-Mark Geusebroek, Arnold Smeulders, Rein van den Boomgaard ISIS, University of Amsterdam

Feature extraction is often based on local image structure like (local) color histograms, or differential geometry. We consider the integration of color and spatial information in a combined spatial color model. The model is based on the Gaussian scale-space theory, establishing well-posed differentiation in the spatio-spectral energy distribution. Hence, the well-known differential invariants from gray level images, as Harris operator, may be extended to the domain of color images. Further we derive photometric invariants which can be measured under different imaging conditions. For each imaging condition results in a different set of invariants. We show the sets to be orderable on degree of invariance

and demonstrate that a higher level of invariance results in less discriminative power between colors. The proposed Gaussian color model is well founded in physics as well as measurement science. Hence, the model is well suited for extracting features as shape, color and texture for content-based image retrieval.

Information Retrieval Methods for Multimedia Objects
Norbert Führ, University of Dortmund, Germany

We discuss concepts of a logic-based approach to multimedia information retrieval. Queries may address four different views on multimedia objects: attributes, logical structure, layout and content. For content representation, one can distinguish between the syntactic, semantic and pragmatic level. Multimedia retrieval should be based on uncertain inference in combination with predicate logic, where vague predicates can be used with predicate logic, for different types of similarity (e.g. color, texture and shape in the case of images). For the hierarchical logical structure, we introduce the concept of augmentation in order to retrieve the most specific sub-tree fulfilling the query. Finally, content-based queries should be preserved using an open world assumption whereas conditions referring to object attributes should employ the closed world assumption.
The DOLORES system developed at the University of Dartmund integrates the concepts described above.

Finding Salient Regions in Images - Density Based Clustering for Segmentation
Eric J. Pauwels, ESAT, K. U. Leuven, and CWI, Amsterdam

In this contribution we argue that Content-Based Image Retrieval can be applied as a tool for automatic keyword propagation and annotation. The idea is that prior visual knowledge is coded as a database of annotated images. When the system is presented with a new and unknown image, it will try and find images (or image regions) in the annotated database that are similar. By looking at the corresponding annotation it then can generate keywords that have a high probability of being relevant for the image under searching.
However, to be able to do this, global comparisons of images are too coarse to be used: one needs to be able to segment the image into regions that can be processed and compared separately. To do this we introduce a non-parametric clustering algorithm that can be applied to different feature spaces. Basically, the procedure looks for the simplest (i.e. smoothest) density that is still comparable with the data. Compatibility is given a precise meaning in terms of the Kolmogorov-Smirnov statistic. The feasibility of this approach is shown in a number of experiments on color and texture segmentation.

Using Invariants for Content-Based Image Retrieval Fundamentals and Apllications, Part II
Sven Siggelkow, Universität Feiburg, Germany

Invariant features remain unchanged when the data is transformed according to a group action. This property can be useful for applications in which a geometric transformation like a change in an object's position and orientation is irrelevant and one is only interested in its intristic features. We discuss features for the group of Euclidean motion (both per gray valued 2D and 3D data) that are calculated by an integration over the transformation group (Haar integrals). The features obtained this way must be modified to meet the requirements of image retrieval. As a result we derive histograms of local appearance features that preserve more local information than the features obtained by global averaging. However the computation time of the features, although it's linear complexity, is too high for applications that either require a fast response (e.g. image retrieval) or have big data set sizes (e.g. 3D tomographic data). So we estimate the features via a Monte Carlo method instead of carrying out a deterministic computation, thus obtaining constant complexity independent from the data set size. For a typical image retrieval application this results in a speedup factor of 100 being able to compute the features in real time. Error bounds for the method are theoretically derived and experimentally verified. We use the features for image retrieval based on a syntactical visual similarity level. They show to be robust to occlusion and also support query by image parts.

Shape Similarity Measure Based on Correspondence of Visual Parts
Longin Jan Latecki, Univ. of Hamburg

A shape similarity measure based on a best possible correspondence of visual parts is presented. Before it is applied, the polygonal contours are simplified by a new process of a discrete curve evolution.
The robust and cognitively adequate performance of this system is confirmed by the excellent result obtained in the Core Experiment Shape-1 in the context of the forthcoming MPEG-7 standard.

Video Abstraction Based on Asymmetric Similarity Values
Sorin M. Jacob, Delft University of Technology

The present work describes a way for extracting most representative key frames from a given set of input shot frames. In order to achieve a low computational complexity, we use a similarity measure based on color histogram distances. However, since global histograms do not provide enough information, we included quantitative information on the local structure of images. The similarity values obtained this way do not enjoy metric properties. Therefore using such values will induce the hierarchical structure of a

multilevel video summary. For this purpose we developed two algorithms. The first one constructs a graph whose nodes are the input key frames and the oriented edges are weighted with the similarity values computed for each image with respect to each other. The second algorithm transforms this graph into a forest of two-level trees, which will generate the two-level hierarchy of key frames.

Data mining in Multimedia Databases
Martin Ester, Hans-Peter Kriegel, Universität München

Clustering and classification are two major data mining tasks, which are also important in the context of multimedia databases. Density-based clustering algorithms require that the number of points in a specified neighborhood exceeds a given threshold. We present an algorithm and discuss different applications of density based clustering, including the generation of thematic maps. Nearest neighbor classification is a simple but effective method. After an introduction to the basic concepts we report some experiences from an application to the analysis of Astronomy images.
In the last part of the talk, we introduce the notion of multiple neighborhood queries that allow a database management system to efficiently support many data mining algorithms. Different techniques of speeding up such sets of queries are presented together with an experimental evaluation of their efficiency on two multimedia databases.

Shape-Based Image Retrieval and Indexing
Remco Veltkamp, Utrecht University, The Netherlands

We approach the retrieval of images in a shape-based way, i.e. we, match and index images on the basis of shapes within the images. We will discuss three issues in shape image retrieval:

- Shape representation, i.e. with which geometric features a shape is represented.
- Similarity measure, i.e. how to shape representations are compared for matching.
- Indexing i.e. the method of access to a large collection of shape representation.

Two shapes ‘match’ if a sequence of translations, rotations, and/or scalings can be found that transforms one shape to the other within a certain error bound. When two shapes do not perfectly match, similarity can be judged by using measures that compare the shapes. Because of occlusions, noise, segmentation faults etc., the similarity measures should be robust and allow partial matching. We will also discuss an indexing scheme that allows fast access of large amounts of shape features, possibly combined with other features such as color.

Similarity Measures for Texture and Grouping

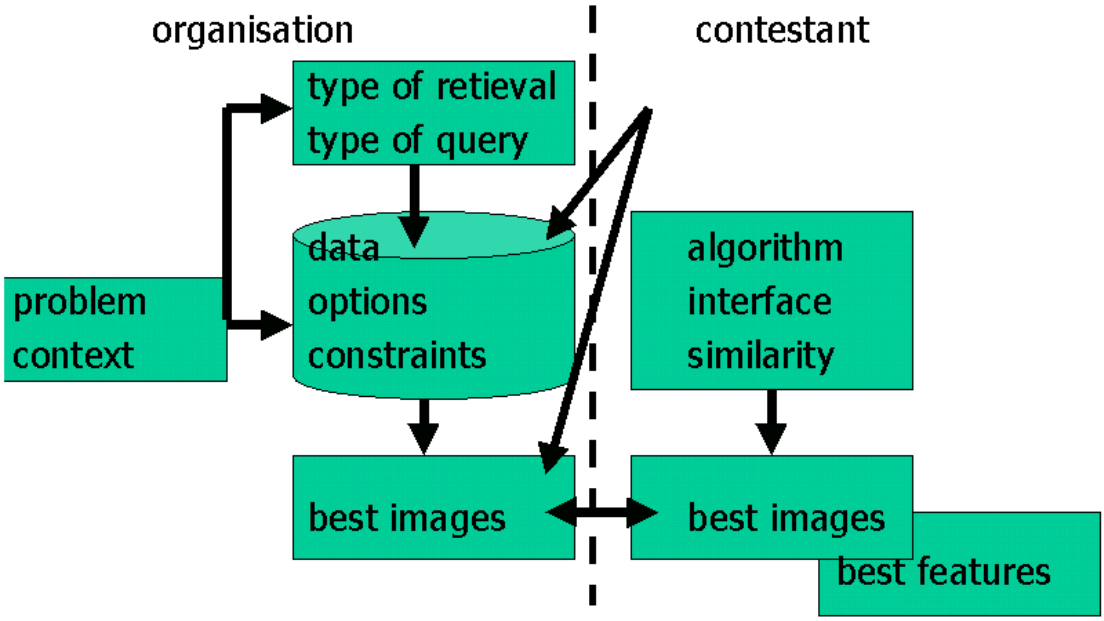
Jan Puzicha, UC Berkeley

Grouping plays an important role in assessing image similarity as is done in image retrieval. The Berkeley Blobworld system is a nice example. In this talk, I introduce a novel image segmentation scheme based on histograms over some feature space. It is based on a statistical mixture model and segmentation is achieved by statistical inference. We provide results for textured image segmentation. Moreover we give an extension to color images, where digital halftoning serves as a method to overcome contouring.

Discussion session Quality Assessment
Alberto del Bimbo, Arnold Smeulders, Remco Veltkamp

Comparing content-based image and video retrieval systems is difficult for various reasons. Many systems are research systems that focus on one aspect of retrieval. Efficiency is hard to compare because they run on different platforms. Effectiveness is hard to compare because they use image databases of different content and size.
In order to initiate ways of comparing retrieval systems, we proposed to organize an image retrieval competition. The envisioned framework is illustrated below. The competition should be held in a number of different categories, for example images of man-made objects, textile, graphics, and art. The competition could be for whole systems, isolated algorithms, etc. The suggested competition rules include:

- The contest is supervised by an international committee
- The participant receives the database at least one month in advance, plus some example queries with corresponding results.
- The contest is performed on site, on a laptop or Web site of the participant.
- The competition is done with new queries which have never been used before.
- The organizers compare the result with the predefined best result, applying some formal evaluation criteria.



The following institutions volunteered to look into the possibilities to collect and make available databases of images:

- Man-made objects: University of Amsterdam
- Textiles: Freiburg University
- Graphics: Utrecht University
- Art: University of Florence