Report to Dagstuhl Seminar 02021, Report No. 329

Content-Based Image and Video Retrieval

J. Malik (UC Berkeley, USA), H.-P. Kriegel (LMU München, Germany), L. Shapiro (Univ. of Washington, USA), R. Veltkamp (Utrecht Univ., The Netherlands)

06.01.-11.01.2002,

Preface

Images and video play a crucial role in Visual Information Systems and Multimedia. There is an extraordinary number of applications of such systems in entertainment, business, art, engineering, and science. Such applications often involve huge collections of images, so that efficient and effective searching for images and video is an important operation.

The previous Dagstuhl Seminar on Content-Based Image and Video Retrieval was the first one on this topic, and turned out to be a big success, as demonstrated by the following two results:

- During the seminar we collectively discussed the problems of performance evaluation and quality assessment of retrieval systems.
- A selection of the presentations has been published as a book in the Kluwer series on Computational Imaging and Vision with the title State-of-the-Art in Content-Based Image and Video Retrieval, Kluwer, 2001.

This motivated us to organize a follow-up seminar, with the central theme "Object recognition for image retrieval". The emphasis of this second seminar will lie on identifying the principal obstacles that hamper progress in content-based retrieval. Fundamental questions such as whether image 'understanding' is necessary for effective image 'retrieval' and whether 'low' level features are sufficient for 'high' level querying. 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.

Topics to be discussed at the seminar include:

Object recognition
Semantic-based retrieval
Indexing schemes
Shape, texture, color, and lay-out matching
Relevance feedback
Visual data modeling
MPEG7 and JPEG2000 issues
Retrieval system architectures

Image and video databases
Feature recognition
Visualizing pictorial information
Video segmentation
Picture representation
Query processing
Perception issues
Searching the web
Delivery of visual information
Benchmarking
Application areas of image and video retrieval

The purpose of this seminar is 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.

Hans-Peter Kriegel, Universität München
Jitendra Malik, University of California, Berkeley
Linda Shapiro, University of Washington
Remco Veltkamp, Utrecht University

Content-Based Image Retrieval in Geneva: Past, Present and Future

Stephane Marchand-Maillet, University of Geneva, Switzerland

Content-based Image Retrieval is a complex field involving various aspects. We present here the developments made since 5 years at the University of Geneva. In particular, our work has led to the release of the GIFT platform for CBIR. The emphasis is placed on a flexible architecture allowing transparent extensions in various directions. To quantify the performances of this basic system, we have investigated the way of performing objective *evaluation*. In particular, we are actively participating in the development of the Benchathlon framework.

We also present extensions and continuation of our work in the direction of *learning* for semantic feature simulation, *multimedia* (and *multimodal*) processing. One important side application we define is multimedia document *annotation* that we think will be crucial in many tasks such as learning and evaluation.

Human image perception and shape retrieval

John P. Eakins, University of Northumbria at Newcastle

Shape retrieval still remains an intractable problem. Through projects such as ARTISAN and SPIRIT we have tried to tackle this problem by developing retrieval techniques based on models of human shape perception. Our prototype ARTISAN shape retrieval systems have achieved some measure of success through implementing rules based on Gestalt principles to group components into perceptually significant regions for matching. Analysis of retrieval failures has led us to propose new matching techniques based on multiple views of an image. Possible ways of implementing these techniques are discussed.

Performance Evaluation in Probabilistic Information Retrieval: The role of generality

D.P. Huijsmans, Leiden University, The Netherlands

Performance Evaluation in Content-based Image Retrieval is about how well the ground-truth dichotomy of relevant/irrelevant items is reproduced by the information retrieval system under study. The four system classes of IR system versus ground-truth are: True Positives, True Negatives, False Positives and False Negatives. System performance is incomplete when only Precision-scope or Precision-Recall graphs are used, because the normalized Precision and Recall values only cover 3 of the 4 system performance classes. *Generality* must be added to complete the set of normalized measurements. We advocate a characterization of IR systems based upon the Total Recall Ideal System in which case the Precision-Recall plane in P,R,G 3D space indicates how well the total recall performance is as a function of the logarithm of Generality.

Interactive Content-based Image Retrieval

Raimondo Schettini, ITIM-CNR, Milano, Italy

We have presented the main features of QuickLook², a general-purpose system that combines in a single framework different approaches, usually considered as alternatives, for querying in image databases. We have shown that it is possible to automate the classification of digital

documents and of photographs in semantic categories. We finally discussed the effectiveness of colour image normalization algorithms and show that Retinex-based pre-filtering improves the effectiveness of colour-based retrieval algorithms. A demo of the system is available at http://quicklook.itim.mi.cnr.it.

Invariant representations
A.W.M. Smeulders, ISIS group, University of Amsterdam, The Netherlands

Of the millions of images possibly containing one and the same object, users of a Content-based Image Retrieval systems will always aim for those actually containing the object regardless the scene-accidental alterations. Among those, the viewpoint rotation magnitude of the camera as well as the conditions of the illumination and the mise-on-scene are scene accidental and hence should be removed in an object-intrinsic representation of the image content. We propose tight sets of invariant features to represent the object on the basis of salient features in the image described by sets with known invariant group. We do so for colour images, specifically and store all alternative invariant groups of features in the database in the recognition that the proper choice of invariance depends on the user's intention, the recording as well (for the retained power of discrimination) as on the content of the database. We derive hierarchically ordered sets of invariants and differential invariants, with still good power to discriminate among 1000 colour patches. We show applications of the invariant representation in edge type classification, CBIR as well as tracking in video.

A Pseudo-Metric for Weighted Point Sets
Remco Veltkamp, Utrecht University, The Netherlands

There are situations, for example in the shape description domain, where the individual points in a feature point set have an associated attribute, a weight. A distance function that incorporates this extra information apart from the points' position can be very useful for matching and retrieval. There are two main approaches for this. One is to interpret the point sets as fuzzy sets. However, a distance measure for fuzzy sets that is a metric, invariant under rigid motion, and respects scaling does not exist. The other approach is the Earth Mover's Distance. However, for sets of unequal weights it gives zero distance for arbitrarily different sets, and it does not obey the triangle inequality.
We have derived a distance measure based on weight transportation, that is invariant under rigid motion, respects scaling, and obeys the triangle inequality, so that it can be used for efficient database searching. This pseudo-metric identifies only weight-scaled versions of the same set. We demonstrate its potential use by testing it on a collection of logos, a set of fish contours, and a collection of 3D polyhedral models.

Knowledge Discovery and Similarity Search in Multimedia Databases
Hans-Peter Kriegel, University of Munich, Germany

The major difference between knowledge discovery in relational and in multimedia as well as in spatial databases is that attributes of the neighbours of some object of interest may have an influence on the object itself. Therefore, such data mining algorithms heavily depend on the efficient processing of neighbourhood relations since the neighbours of many objects have to

be investigated in a single run of a typical algorithm. We define a small set of database primitives and we demonstrate that typical data mining algorithms such as clustering, characterization and trend detection are well supported by the proposed database primitives.
In the second part of this talk, we focus on similarity search in multimedia databases which is highly application- and user-dependent. Therefore, we derive similarity models to be adaptable to application specific requirements and individual user preferences. Examples include flexible pixel-based shape similarity, 3D shape histograms and quadratic forms, resulting in ellipsoid queries in high-dimensional data spaces. The talk concludes showing some snapshots of our similarity search system.

Material Recognition for Content Based Image Retrieval
Jan-Mark Geusebroek, University of Amsterdam, The Netherlands

One of the open problems in Content-based Image Retrieval is the recognition of materials present in an image. Knowledge about the set of materials present gives important semantic information about the scene under consideration. For example, detecting sand, sky and water certainly classifies the image as beach.
We try to tackle the problem of material recognition in two stages. First, the material reflectance characterized by invariant colour properties distinguish matte materials from glossy ones like metals. Comparison of the spatial response of various invariants leads to such a characterization. Secondly, the “touch” or roughness of a material may be characterized by investigating physical invariant texture properties. Therefore, we study the propagation of transformation groups through the Gaussian NJet. We demonstrate the NJet to characterize the image as points in a high-dimensional scatter plot of the NJet components. Characterization of materials is then based on point cloud matching with prototype Njets. Finally, matching can be improved by fitting the NJet cloud to a statistical distribution. We show the image derivatives to obey a symmetric Weibull distribution, where the shape parameter varies between an exponential distribution and a Gaussian distribution. Matching the parameters of the Weibull distribution may lead to material recognition.

Content-based Image and Video Retrieval: New (and last) Avatar of Digital Picture Processing?
Jean-Michel Jolien, INSA Lyon, France

We present during this seminar some thoughts on the emerging domain of Visual Information Management. We set up these thoughts in a historical analysis of the unsolved questions of the past and mainly the assumption/need of an objective representation of the image domain. We try to show what is similar to the problems we had to face in the past and what is really new in the VIM domain. We set up some trends toward a better understanding of the key approaches and propose a personal view. Then we focus on query by example. In this particular case, the user provides the system with an image. A classic way of optimising the system’s answer is the so-called relevant feedback technique. The user is asked to score the images returned by the system. We propose that we first do some kind of a priori (and not a posteriori) relevant feedback on a set of transformations of the query image. This allows the system to learn what is important in the query. This follows the statement of Molles and Rhomer, who argued that any image set has its own image’s theory. The system must understand the “theory” behind the query. We give an outline of how this a priori relevant feedback can be done with basic image transforms.

Object Recognition for Content-Based Image Retrieval
Linda M. Shapiro, University of Washington, USA

The standard paradigm in first generation content-based image retrieval systems has been the query-by-example paradigm. The user presents a query image to the system which retrieves the most similar images from the database. This paradigm, while useful for some very specific applications such as medical images or trademark images, is not at all suitable for the general applications of finding images from a varied database for use in marketing, advertising, or other literature that will illustrate a point. Users of such a system do not have such an image, only a description of what they want, usually in terms of words involving objects and concepts. Our work is an attempt to provide automatic indexing of images through recognition of objects and concepts. We have developed algorithms for recognizing such objects as boats, vehicles, and buildings. The features for these and other common objects should be accumulated and used to train the system to recognize additional objects. A hierarchical classifier system is proposed for this difficult learning task.

Retrieving Scenes in Videos
Andrew Zisserman, University of Oxford, GB

We discuss the problem of determining whether two images are perspective images of the same 3D scene or not. This question is more difficult to answer when the camera viewpoints differ substantially because of the change in the visual appearance of the scene between images: surface foreshortening differs, there may be partial occlusions, there may be lighting changes.

The approach taken is to determine local scene descriptor vectors which are invariant to 2D geometric transformations and 1D photometric transformations. These descriptors are computed independently in each image, and there may be 1000's of such descriptors per image. Of particular importance is that the shape of the descriptor adapts to cover the same scene region covariantly with viewpoint. The descriptors are matched between images based on the vector of invariants. Two images are deemed matched if sufficient of these correspondences are consistent with an epipolar geometry relationship between the images. The method is applied to scene matching in videos, where shots of the same scene are identified by matching key frames.

Extraction of Artificial text for Semantic Indexing
Christian Wolf and Jean-Michel Jolion, INSA Lyon, France

The systems currently available for content-based image and video retrieval work without semantic knowledge, i.e. they use image processing methods to extract low level features of the data. The similarity obtained by these approaches does not always correspond to the similarity a human user would expect. A way to include more semantic knowledge into the indexing process is to use the text included in the images and video sequences. It is rich in information but easy to use, e.g. by keyword based queries. In this talk we present an algorithm to localize artificial text in images and videos using a measure of accumulated gradients and morphological post-processing. The quality of the localized text is improved by robust multiple frame integration. A new technique for the binarization of the text boxes is proposed. Detection results and OCR results for commercial OCR software are presented.

The second part of the presentation summarizes open problems which need to be tackled in the future, especially the detection of moving scene text. The current state of the art is given and a possible roadmap including possibilities how to improve the models is given.

Exploiting Text and Image Feature Co-occurrence Statistics in Large Datasets

Kobus Barnard, University of California, Berkeley, USA

I describe recent work on two related problems: improving access to large image datasets and exploiting them for computer vision. The problems are related because providing effective access to such collections requires representing image semantics. For example, a user searching for a tiger image will not be satisfied with an image with a plausible histogram – tiger *semantics* are required. To capture such semantics requires models using both text associated with the images and features computed for image regions. Large amounts of data suitable for this approach are available (Corel: 40000 images, museum data video with speech recognition, web images). The learned model supports browsing, searching based on text, region features, or both, as well as novel applications such as suggesting images for illustrating text passages (auto-illustrate), attaching words to images (auto-annotate), and attaching words to *specific* image regions (recognition).

Repeated Pattern Reception using Symmetric Groups

Yanxi Liu, Carnegie Mellon University, Pittsburgh, USA

Humans have an innate ability to perceive symmetry. It is not obvious how to automate this powerful insight. While there has existed a mature mathematical theory on periodic patterns for over a century, namely the theory of crystallographic groups. Our work appears to be the first to automate the process of classifying digitised frieze and wallpaper patterns into their respective symmetry groups, and generate their representative motifs.

Even though the appearance of a periodic pattern can change drastically and infinitely under affine transformations, its symmetry groups stays finite and stable, and thus provides a good index for regular patterns (textures) viewed at arbitrary angles. A complete and concise skewed symmetry group "migration map" is constructed for the 17 wallpaper groups, that intertwine relationship among the symmetry groups with small, distinct orbits. Besides regular pattern indexing and retrieval, the applications of this computational model also include: orientation estimation, image compression, texture synthesis, texture replacement, and gait analysis.

Classification-driven Feature Space Reduction for Semantic-based Biomedical Image Indexing and Retrieval

Yanxi Liu, Carnegie Mellon University, Pittsburgh, USA

Biomedical image databases provide challenging yet quantifiable testbeds for semantic-based image indexing and retrieval. General image features of diverse 2D images (e.g. colour, texture) used by existing content-based retrieval systems often fail to be effective discriminators for biomedical image sets with subtle, domain-specific differences. Furthermore these measures do not necessarily reflect the meaning of an image.

We are exploiting novel image features for volumetric biomedical images, that have predefined semantics. Even though each individual feature can be weak or fallible, the most discriminative feature subset(s) can be found from a large set of potential indexing features using classification-driven feature selection methods. Our goal is to simultaneously minimize the feature space dimension and increase the discriminating power at retrieval time. We show the feasibility of our approach through 3 applications:

1. Statistical brain asymmetry measures on pathological neuro-images for pathology discrimination.

2. Facial asymmetry measures on facial expression videos for human identification.
3. Geometry/shape features combined with local texture features on multispectral pap smear images for cancer cell detection.

On Invariants in Image Retrieval
Hans Burkhardt, University of Freiburg, Germany

The talk motivates the use of invariants for image retrieval purposes. An equivalence relation between images is defined on the basis of geometric transformations like individual independent Euclidian motion if objects in a scene. Three canonical methods to construct invariants are described: group averaging (Haar integrals), Lie theory (differential approach) and normalization techniques. The talk concentrates on Haar integrals with monomial functions of local support. Invariance is demonstrated for global Euclidian Motion, individual Euclidian Motion, articulated objects and for topological deformations. Examples are given for image retrieval of planar patches and an extension to 3D invariant recognition of objects is demonstrated (a project for the recognition of pollen).

The Informedia Digital Video Library System and the TREC Video Retrieval Evaluation
Alexander Hauptmann, Carnegie Mellon University, Pittsburgh, USA

The Informedia Digital Video Library Project has demonstrated the successful application of speech, image and natural language processing on high quality broadcast television material in automatically creating a rich, indexed, searchable multimedia information resource. Over 2 Terabytes of data have been collected since 1997 and automatically extracted metadata has been generated. The metadata includes shotbreaks, speech transcriptions, titles, topics, detected faces, detected and transcribed text captions in the video image, named entities, especially locations and non-speech audio-characteristics. The talk illustrates how this imperfectly extracted metadata can be indexed and searched in an integrated interface.
The second part of the talk focused on the 2001 TREC evaluations for Video retrieval systems. Participants were given a text description of the information need (*query*) as well as possibly examples of images, video clips or audio segments. Eleven hours of video were searched for shots that contained possible relevant video. While performance of the interactive Informedia system was quite good, the automatic systems, both from CMU as well as other participants, exhibited only minimal success. This indicates the difficulty of the task and underscores the fact that a variety of components (speech recognition, speaker identification, video OCR, face detection and matching, and, of course, similarity matching) need to become better as individual components, and also need to be well integrated to succeed in the video retrieval task.

Feature Histograms for Content-based Image Retrieval
Sven Siggelkow, University of Freiburg, Germany

In many applications the absolute object position and orientation are irrelevant. So we start from construction invariant features, which only consider the remaining object characteristics. Based on a general construction method we derive invariant feature histograms that catch different cues of image content: features strongly influenced by colour (joint histograms of

monomial kernel functions calculated on the colour layers) and textual features that are robust to illumination changes (kind of a rotation-invariant fuzzyfied Local Binary Patterns). For improving clustering properties, we remove discontinuities in the histogram by using fuzzy histograms which are related to kernel-based probability density estimators. In order to speed up the feature calculation we apply a Monte-Carlo estimation of the features rather than a complete calculation, but we are still able to predict the error depending on the number of samples.
Upon these theoretical considerations, we built two image retrieval systems: The first one presented, SIMBA (Search Images By Appearance), deals with natural images of general content. By weighting colour against texture, the user can adapt the query according to his needs or the character of the image, The second system, MICHELScope, considers looking up stamps from a database of 13000 stamps. In addition it can be used to find series of stamps, which share some common characteristics. As presented in the live-demo, the features are robust to quite much variation of the stamps motives.

Interactive Image Retrieval in Specific and Generic Image Databases
Nozha Boujemaa, INRIA, Rocquencourt, France

For designing an effective image retrieval systems, we find it convenient to divide image databases in two categories:

- the first category concerns specific image databases, for which a ground truth is available. When indexing the database, maximizing the system efficiency. We have developed specific signature for face recognition and detection, fingerprint identification.
- The second category includes databases with heterogeneous images where no ground truth is available or obvious. Examples include stock photography and the WWW. The user should be assumed to be an average user (not an expert). In this context, generic image signatures are computed in order to describe general visual appearance such as colour and texture. We present the weighted histogram signatures for integrated colour/texture information.

Results and examples were presented by our CBIR IKONA which has a client/server architecture. Precise search by local descriptions and query method was presented. It is based on colour point of interest. Applications with criminal investigation department was shown.
Other results were presented on image database overview by clustering method as well as cross-media indexing method. In the latter case, keyword propagation was performed based on visual similarity.

Database Support for Content-Based Retrieval
Thomas Seidel, University of Konstanz, Germany

For many application domains, similarity search is a quite subjective task for which the user's preferences should be taken into account. The talk addresses the following aspects:

(1) Geometric Similarity. 3D shape histograms are able to model the similarity of extended spatial objects including protein structures from biomolecular databases or mechanical parts from CAD databases. Quadratic form distance functions help to incorporate the user's notion of similarity in mind and to cope with small displacements of the shapes.

(2) Relevance Feedback. An important approach to take the user's needs into account is to iteratively refine the queries by relevance feedback. When considering the cross-

correlations of the feature dimensions for the positively marked answers, the query engine has to support quadratic form distance functions with varying correlation matrices.

(3) Ellipsoid Queries. For both concepts, geometric similarity as well as relevance feedback, the quadratic form distance functions yield elliptical query regions. On top of our previous solutions for dynamic ellipsoid queries, we present new approximations that apply to recent vector quantization techniques for indexing high-dimensional feature vectors.

Image retrieval in the presence of important viewpoint changes and with automatically constructed models

Cordelia Schmid, INRIA Rhone-Alpes, France

In this presentation we address two aspects of image retrieval. First, we present the retrieval of an object or a scene in the presence of important viewpoint changes (scale changes and changes in viewing angle). The approach is based on the detection of affine invariant interest points. These points are used to characterize the image; the affine transformation associated with each point allows to compute affine invariant descriptors. Experimental results for retrieval show an excellent performance up to a scale factor of 4 and important changes in viewing angle for a database with more than 5000 images.

Secondly, we automatically construct visual models for the retrieval of similar images. Models are constructed from a set of positive and negative sample images where no manual extraction of significant objects or features is required. Our model allows to efficiently capture “texture-like” structure and is based on two layers: “generic” descriptors and statistical spatial constraints. The selection of distinctive structure increases the performance of the model. Experimental results show a very good performance for retrieval as well as localization.