04021 Abstracts Collection
Content-Based Retrieval
— Dagstuhl Seminar —

Jitendra Malik[1], Hanan Samet[2], Remco Veltkamp[3] and Andrew Zisserman[4]

[1] UC Berkeley, US
[2] Univ. of Maryland, US
hjs@umiacs.umd.edu
[3] Utrecht Univ., NL
remco.veltkamp@cs.uu.nl
[4] Univ. of Oxford, GB
az@robots.ox.ac.uk

Abstract. From 04.01.04 to 09.01.04, the Dagstuhl Seminar 04021 "Content-Based Retrieval" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Object recognition, semantic-based retrieval, indexing schemes, matching algorithms, shape, texture, color, and lay-out matching, relevance feedback

Sports highlight recognition for semantic annotation and video adaptation

Marco Bertini (University of Firenze, I)

Automatic annotation of semantic events allows effective retrieval of video content. In this presentation, solutions for highlights detection in sports videos are shown. The proposed approach exploits the typical structure of a wide class of sports videos, namely those related to sports which are played in delimited venues with playfields of well known geometry, like soccer, basketball, swimming, track and field disciplines, and so on. For these sports, a modeling scheme based on a limited set of visual cues and on finite state machines that encode the temporal evolution of highlights is presented, that is of general applicability to this class of sports. Visual cues encode position and speed information coming from the camera and from the object/athletes that are present in the scene, and

are estimated automatically from the video stream. The semantica annotation obained is then used to select objects and events of interest for the viewer, in order to adapt the video quality according to user preferences. The performance of these systems should be evaluated especially with regards to the user's satisfaction, in terms of viewing quality and cost paid, taking into account his/her preferences about the content. To analyze the user's satisfaction, we propose a new performance measure that explicitly takes into account the user's preferences and considers the number and type of errors produced by the annotation engine and the way in which these errors affect the compressed video quality and bandwidth allocation.

Keywords: Video annotation, automatic annotation, video adaptation, transcoding

Mental Image Search

Nozha Boujemaa (INRIA Rocquencourt, F)

To retrieve images by visual content, the most commonly used Query-by-Example paradigm is not effective since the user usually does not have an example to start the search but rather a mental image. To overcome this problem we propose new image retrieval paradigm: logical composition of region categories. The user can directly specify the logical composition of visual patches; which are parts of his target mental image. After performing clustering in the image regions feature space, the query interface consists of a visual summary of the image regions available in the database which constitutes a visual thesaurus. A new symbolic indexing and querying approach is presented which relates closely to that of text retrieval. This approach could be easily combined with high level semantics labelling and querying. Simple and generic, it is extendable to multimedia document retrieval indexed by other physical content descriptors.

Joint work of: Boujemaa, Nozha; Fauqueur, Julien

Extracting Generic Clusters using Edge-based Features from Image Sequences

Gerd Brunner (Universität Freiburg, D)

We present a method for clustering geometric features and apply it to content-based image retrieval (CBIR) in short image sequences. Straight line segments and their connectivities carry significant information about object shapes. The geometric structure of a scene can be extracted from embedded spatial and angular line segment relationships.

By using hierarchical clustering, generic structure information on different scales are obtained. In a further step line cluster interrelationships are investigated and employed to vehicle search in CBIR.

Comprehensive tests have been performed on an image database consisting of short image sequences from an outdoor surveillance camera.

Joint work of: Brunner, Gerd; Burkhardt, Hans

Qualitative Shape Matching for Object and Action Recognition

Stefan Carlsson (KTH Stockholm, S)

Recognition of object classes poses problems of shape and appearance variation that are much harder than those encountered in the recognition od specific objects. We will present an algorithm for matching perceptually corresponding shapes based on representing order properties of the image geometry. Using ordinal as opposed to metric structure captures qualitative shape properties that typically define object and action classes and it allows for a wider range of variation in the data to be matched. In the first part we will present results in recognition of object and action classes based on an algorithm for computing point to point correspondence between shape prototypes and the image to be recognized. In the second part we will present preliminary results on direct shape matching of images by computing order properties from the image gradient field

An Efficient Indexing and Search Technique for Multimedia Databases

Michael Clausen (Universität Bonn, D)

We present a novel index-based approach for searching multimedia databases by content. Our approach integrates methods from classical full-text retrieval with the mathematical concept of groups acting on sets. This yields a flexible framework applicable to a wide range of content-based search problems such as audio- or image-identification.

We propose space-efficient indexing methods as well as very fast fault-tolerant searching algorithms. In contrast to other approaches, our query response times decrease with incresing query complexity. As a further main benefit, a concept of partial matches is an integral part of our technique.

Several prototypic applications are discussed demonstrating the capabilities of our new technique.

Joint work of: Clausen, Michael; Bardeli, Rolf; Körner, Heiko; Kurth, Frank; Mosig, Axel; Röder, Tido; Ribbrock, Andreas

Recognition of Musical Variations by Harmonic Modelling

Tim Crawford (Goldsmiths College - London, GB)

The concept of generalised musical similarity for information retrieval is well known to be problematic. For certain, simple or reduced types of musical data it is sometimes possible to use a measure very close to text-based 'edit distance', but this is rarely the case with 'real-world' musical data, such as that generated from performances. In the OMRAS (Online Music Recognition and Searching) project we developed a method for approximate matching of polyphonic queries in a database of polyphonic music based on the techniques of language modelling that has proved very suitable for work with corrupt data, being remarkably robust to errors introduced by faulty audio recognition. This involves the probabilistic modelling of the harmonic content of a piece by reference to the 24 major and minor triads, and using a distance measure between them derived from widely-accepted music-perception experimental results to provide a characteristic model for that piece. Searching a database for pieces similar to the query is simply a matter of modelling the query in the same way and ranking each database piece by the divergence of its model from that of the query. While the technique remains still in the early stages of development (for example, no time-based data is yet accounted for in the model) it seems that by way of spinoff it holds much promise for certain kinds of musicological investigation.

Recognising musical variations algorithmically poses special problems. By definition, a variation must be different from its 'theme', yet at some level it retains a recognisable similarity to it. Composers have used a wide variety of techniques for variations, but for a large proportion of the repertory there is a strong tendency to retain the broad harmonic outline, or contour, throughout. While this is most definitely true of variations composed before the late 18th century, it actually holds good for much later music as well, although there are of course many exceptions. Preliminary work on using the OMRAS technique suggests that the harmonic-modelling approach could be a very useful tool for recognising this kind of variation.

Some early results from OMRAS experiments on four controlled query-repertories will be described. These were partly devised as a way of testing the performance of the OMRAS system, but can perhaps also be seen as useful musicological investigations in their own right. The query sets comprise:

- a) Bach's 'Goldberg' variations; treating each variation in turn as a query, how well can we retrieve all the other variations?
  b) 4-part Bach chorales from which 1, 2 and 3 complete single voices have been removed; how many (and which) voices are necessary to retrieve the complete version?
  c) 4-part Bach chorales which have been treated to systematic 'degradation' (by random deletion, insertion and/or transposition of notes); how robust is the system to error?
  d) a set of about 25 different versions of the same lute pavan segmented

into strains A, B and C, some of which also contain more or less elaborate 'divisions' or variations, A', B' and C'; how well can we recognise the plain from the elaborated versions, and vice versa?

In each of these cases, it is as interesting (for the musicologist) to look at high-ranking 'non-relevant' as well as 'relevant' matches; in many cases they suggest (though this needs further testing) that the language-modelling approach may prove useful in style identification.

How good are Current Shape Description Techniques?

John Eakins (University of Northumbria, GB)

Despite over 30 years' research, few techniques for shape similarity retrieval appear to be reliable enough for commercial applications such as trademark registration. One reason for this is the relative scarcity of effective benchmarking studies. Following extensive but unpublished comparative tests, the ISO MPEG-7 standard now incorporates two shape descriptors - the angular radial transform (ART) and curvature scale-space (CSS) coefficients. But whether these are in fact sufficiently effective for commercial use is still open to question.

Our own experiments on trademark image retrieval suggest that, for this application at least, other types of shape descriptor can outperform those recommended in the MPEG-7 standard, with a combination of simple, Fourier, and Rosin descriptors providing the most effective combination. We are currently trying to gain further insight into the relative effectiveness of different combinations of descriptors through principal components analysis and Sammon mapping. Some recent results from these investigations will be displayed at the meeting.

Joint work of: Eakins, John; Edwards, J.D.; Riley, EJ

Image Retrieval by Evaluation of Non-Lienar Kernel Functions around Saliener Points

Alaa Halawani (Universität Freiburg, D)

Feature histograms based on the evaluation of Haar integrals with nonlinear kernel functions were used successfully for the purpose of invariant content based image retrieval. In addition to being invariant to rotation and translation, the features have the advantage of preserving structural information of the image.

The work presented here concentrates on the idea of calculating these features by evaluating the kernel functions around a small set of preselected points. These points are called the salient points and represent, together with their neighborhood, the most important visual information in an image. The use of

these salient points leads to a better representation of the image. Compared to previous work, experiments show that this method gives better retrieval results without introducing extra computational overhead.

Joint work of: Halawani, Alaa; Burkhardt, Hans

Aspects of semantic Video Analysis

Otthein Herzog (Universität Bremen, D)

The large amount and the ubiquitous and increasing availability of multimedia information (e.g., video, audio, image, and including also text documents) requires efficient, effective, and automatic annotation and retrieval methods. As videos play a major role in multimedia content-based analysis, retrieval of videos becomes an issue. It is sensible, too, to capitalize on the work done in Information Retrieval, and work on an integrated methodology for all types of multimedia documents.

Automatic analysis and annotation of video documents covers a wide range of methods to extract information on different levels of abstraction. First, there are syntactical approaches like e.g., the detection shot boundaries or the extraction and transcription of text inserts. Inserts are very helpful in order to extract content of a video, automatically. This task is called Video OCR. While there are numerous approaches on text/background separation, less work has been done on a fast detection of the relevant frame intervals within the entire video. Therefore, we present a new approach on real-time detection of frame intervals containing textual inserts.

A step further in automatic annotation is the analysis on a semantical level like e.g., object recognition as well as scene classification and interpretation. Another aspect of our actual work which we present in this work is the spatio-temporal analysis of dynamic scenes. Based on a qualitative description of object motion, dynamic scenes are interpreted on a semantic level and upcoming motion situations are predicted.

Joint work of: Herzog, Otthein; Hermes, Thorsten; Miene, Andrea; Wilkens, Norbert

Image indexing using strings

Jean-Michel Jolion (INSA - Lyon, F)

The purpose of this talk is to introduce a new representation scheme for images based on strings. Many works in the image retrieval domain have already used interest points as the basis for image representation but the spatial structure of these points as well as their relevance are mainly used (if used) as one more parameter. We propose to set an image representation as a list (so with a given

order) of particular points extracted from an image. Then we have also to focus on the valuation of this list, i.e. the information we attached with each element of this list. In many applications based on local descriptors, a vector of differential invariants or equivalent measures are extracted. We introduce a valuation based on local contrast by means of symbolic patterns.

This new representation scheme is so based on ordered list of symbolic patterns, i.e. a string.

Comparison of the representation makes use of recent results on statistical processing of sets of strings.

The proposed talk will focus on principles of this approach as well as the general architecture for extraction and comparison of strings.

Some preliminary results will illustrate the talk.

Joint work of: Jolion, Jean-Michel; Simand, Isabelle

Studying digital imagery of ancient paintings by mixtures of stochastic models

Jia Li (Penn State University, USA)

This talk will address learning based characterization of fine art painting styles. The research has the potential to provide a powerful tool to art historians for studying connections among artists or periods in the history of art. Depending on specific applications, paintings can be categorized in different ways. In this paper, we focus on comparing the painting styles of artists. To profile the style of an artist, a mixture of stochastic models is estimated using training images. The 2-D multiresolution hidden Markov model (MHMM) is used in the experiment.

These models form an artist's distinct digital signature. For certain types of paintings, only strokes provide reliable information to distinguish artists. Chinese ink paintings are a prime example of the above phenomenon; they do not have colors or even tones. The 2-D MHMM analyzes relatively large regions in an image, which in turn makes it more likely to capture properties of the painting strokes. The mixtures of 2-D MHMMs established for artists can be further used to classify paintings and compare paintings or artists. We implemented and tested the system using high-resolution digital photographs of some of China's most renowned artists. Experiments have demonstrated good potential of our approach in automatic analysis of paintings. Our work can be applied to other domains.

Joint work of: Li, Jia; Wang, James Z.; Sate, Penn

Large Visual Document Collections

Stéphane Marchand-Maillet (CUI - Geneva, CH)

Most current Multimedia Information Management frameworks are query-based. That is, they mostly rely on the assumption that the user has a good idea of what (s)he is looking for. This can be mapped onto the concept of Query-by-Example, where the user is able to produce an example of what (s)he is looking for. Browsing is another technique for searching information. It also assumes that the users holds the definition of a specific target. In both concepts, the user should be able to produce a query describing the information need.

Here, we look at information management in a "queryless" context. The user is faced with a (large) number of multimedia data. The tools we wish to describe here should help the user getting a comprehensive view of the content of these collections effectively. Our baseline is a simple view system that shows items in random order.

We use common techniques of statistical clustering and dimension reduction to reach our goal. We have also developed an original approach for characterising information summaries as solutions of global optimisation problems. Once optimal subset are found, we are left with the problems of visualising accurately our data. We propose and discuss several solutions to this problem.

Comparing Text and Shape Matching for Retrieval of Online 3D Models

Patrick Min (Utrecht University, NL)

While retrieval of 3D models from the Web has become a focus of research over the last few years, an important question has largely been overlooked: How do shape-based and text-based retrieval methods compare? Previous work has shown that textual queries of captions are useful for searching image databases (e.g., Google Image Search, and [Sable03, Smith96]) – how about for 3D models?

To investigate this question, we ran classification tests using text-based and shape-based matching methods on a large database of 3D models downloaded from the Web [Shilane04]. We tested seven different text classification methods (e.g., TF/IDF, Naive Bayes, etc.), matching textual descriptors for each 3D model consisting of words from all possible combinations of eight different sources (e.g., its filename, identifiers found within the file, the web page context, etc.). We also investigated using Wordnet (a lexical database) to add synonyms and hypernyms (category descriptors) to the textual representation [Miller95].

Nevertheless, we found that shape-based matching outperforms text-based matching in all our experiments. The main reason is that 3D models found on the Web are poorly annotated – they generally appear in lists on web pages, annotated only with cryptic filenames or thumbnail images. The text sources that proved to be most useful were text from the model file itself and the Wordnet

synonyms and hypernyms of the filename and link text. Interestingly, we found that combining the results from the text and shape matching methods improves performance over either alone, which suggests that further research is warranted on search methods that consider multiple attributes of 3D models.

Geometric Approach to Music Retrieval

Veli Mäkinen (University of Helsinki, FIN)

We represent music symbolically as sets of points or sets of horizontal line segments in Euclidean plane. Via this geometric representation we cast transposition invariant content-based music retrieval (CBMR) problems as ones of matching sets of points or sets of horizontal line segments in plane under translations. For finding the exact occurrences of a point set (the query pattern) of size m within another point set (representing the database) of size n, we give an algorithm with running time $O(mn)$. Moreover, we give two $O(mn \log m)$ algorithms first of which finds partial occurrences of a given point set, and the other the largest overlap between the line segments representing a query pattern and a database. Some experimental results on the performance of the algorithms are reported.

Joint work of: Ukkonen, Esko; Lemström, Kjell; Mäkinen, Veli

Content-based Image Retrieval and Pictorial Query Specification Using Fourier Descriptors on a Logo Database

Hanan Samet (University of Maryland - College Park, USA)

A system that enables the pictorial specification of queries in an image database is described. The queries are comprised of rectangle, polygon, ellipse, and B-spline shapes. The queries specify which shapes should appear in the target image as well as spatial constraints on the distance between them and their relative position.

The retrieval process makes use of an abstraction of the contour of the shape which is invariant against translation, scale, rotation, and starting point that is based on the use of Fourier descriptors.

These abstractions are used in a system to locate logos in an image database. The utility of this approach is illustrated using some sample queries.

Joint work of: Samet, Hanan; Folkers, Andre; Soffer, Aya

3D-Model Retrieval

Dietmar Saupe (Universität Konstanz, D)

Content-based similarity search is an important and challenging topic in digital libraries research. In this talk, we focus on improving the effectiveness of similarity search in 3D object repositories from a system-oriented perspective. Motivated by an effectiveness evaluation of several individual 3D retrieval methods, we research a selection heuristic, called purity, for choosing retrieval methods based on query dependent characteristics. We show that the purity selection method signiffcantly improves the search effectiveness compared to the best single methods. We then demonstrate that retrieval effectiveness can be further boosted by considering combinations of multiple retrieval methods to perform the search. We propose to use a dynamically weighted combination of feature vectors based on the purity concept, and we experimentally show that the search effectiveness of our combined methods by far exceeds the effectiveness of our best implemented single method.

Keywords: Similarity search, 3D models, spherical harmonics, impurity, Karhunen Loeve transformation

Using Transportation Distances for Measuring Melodic Similarity

Rainer Typke (Utrecht University, NL)

Many existing methods for measuring melodic similarity use one-dimensional textual representations of music notation, so that melodic similarity can be measured by calculating editing distances.

We view notes as weighted points in a two-dimensional space, with the coordinates of the points reflecting the pitch and onset time of notes and the weights of points depending on the corresponding notes' duration and importance. This enables us to measure similarity by using the Earth Mover's Distance (EMD) and the Proportional Transportation Distance (PTD), a pseudo-metric for weighted point sets which is based on the EMD.

A comparison of our experiment results with earlier work involving the same data shows that by using weighted point sets and the EMD/PTD instead of Howard's method (1998) using the DARMS encoding for determining melodic similarity, it is possible to group together about twice as many known occurrences of a melody within the RISM A/II collection. Also, the percentage of successfully identified authors of anonymous incipits can almost be doubled by comparing weighted point sets instead of looking for identical representations in Plaine & Easie encoding as Schlichte did in 1990.

Unlike string-based methods, transportation distances are well suited to working with polyphonic music.

Topics of this talk:

- Using transportation distances for comparing melodies, both monophonic and polyphonic.
- Indexing: Vantage point method
- The prototype of a search engine using these methods.

Joint work of: Typke, Rainer; Giannopoulos, Panos; Veltkamp, Remco C.; Wiering, Frans; van Oostrum, René

Part-Based Retrieval

Remco Veltkamp (Utrecht University, NL)

Effective and efficient shape-based image retrieval is a non-trivial task.

Effectiveness needs the abilibty to perform partial matching, the matching of only a part of an object with part of another object. However, partial matching often implies that the underlying similarity measure does not satisfy the triangle inequality. This again implies that indexing systems that make use of this property cannot by used to make the retrieval efficient.

Our approach is to decompose the shape into parts first, and then to match parts with a similarity measure that does satisfy the triangle inequality.

Decomposition of a shape makes sense, because the human visual system uses a part-based representation.

We propose a novel type of decomposition for polygonal shapes. Our method is the first one that is based on the straight line skeleton. Compared to the medial axis, the straight line skeleton has a few advantages: it contains only straight segments and has a lower combinatorial complexity. The decomposition is invariant to rigid motions and uniform scalings. We present results indicating that it provides natural decompositions for a variety of shapes, and show a part-based retrieval application based on this decomposition.

Keywords: Shape matching, partial matching

Joint work of: Veltkamp, Remco; Tanase, Mirela

From Content-based Image Retrieval to Automatic Linguistic Indexing of Pictures

James Wang (Penn State University, USA)

The need for efficient content-based image retrieval has increased tremendously in many application areas such as biomedicine, military, commerce, education, and Web image classification and searching. In this talk, we present our research in the area of intelligent image indexing and retrieval. We developed a wavelet-based approach for feature extraction and an integrated region matching (IRM) technique for matching region features. An image in the database is represented

by a set of regions, roughly corresponding to objects, which are characterized by color, texture, shape, and location. A measure for the overall similarity between images is developed as a region-matching scheme that integrates properties of all the regions in the images. Our recent research focuses on developing ALIP (Automatic Linguistic Indexing of Pictures), a system to index images using automatically learned statistical models. Categorized images are used to train a dictionary of hundreds of concepts automatically based on statistical modeling. Images of any given concept category are regarded as instances of a stochastic process that characterizes the category. To measure the extent of association between an image and the textual description of a category of images, the likelihood of the occurrence of the image based on the stochastic process derived from the category is computed. A high likelihood indicates a strong association. Experiments show that the system has high potential in linguistic indexing of images. In the talk, we will also cover some of our most recent work. More information is available at http://wang.ist.psu.edu

Why music retrieval on symbolic data?

Frans Wiering (Utrecht University, NL)

In the Orpheus project we use transportation distances as a similarity measure in retrieval from encoded music notation. Rainer Typke will focus on the actual methods we are using in his talk; I will discuss two underlying issues:

1 . why retrieval on symbolic data (encoded music notation) rather than digital audio;
2. what is musical similarity.

1. At first sight, audio seems to be more suitable for music retrieval. However, for broad categories of queries features are needed that are hard to extract. Such extraction is much researched, and it is likely that the resulting representation will have a great deal of resemblance to music notation. So in a way, we are starting from the other end.

Symbolic data has its problems too, including poor availability. So far, searching is usually based on textual representations. These work best when music is reduced to one dimension (usually pitch) which means that rhythm and polyphony are generally neglected.

Fundamental question here is: are notes what people perceive and remember?

2. What is musical similarity? Musical similarity comes in many flavours, from very generic similarity (rock 'n' roll, classic) to very specific (repeated pieces of melody in one piece of music). We are looking at the specific end, i.e. fragments of music that may be regarded as instantiations of the same 'archetype'.

It is not hard to find theories about musical similarity. Music theory itself is a discipline that prides itself on going back to the ancient Greeks, and, early middle ages excepted, there is no lack of sources. What these share is speculation

(often quite insightful); what is notably lacking is two things: consensus and experimental verification. It is also strongly directed towards unification: we have so much powerful theory at hand that we can show that anything is somehow related to everything else, but have no equally developed mechanism that tells us where to stop. It is also very much focussed on notation, not on what people perceive and remember. This again leads to the question: are notes what people perceive and remember? (Partly tentative) answers come from music cognition. Melodies, for example, seem to be remembered mainly by contour rather than by notes. We hear chords rather than clusters of notes (this has been long acknowledged in music theory), and reduce seemingly quite different melodies to the same chord progression.

We are currently examining how knowledge about music cognition can be integrated in our search engine. For example, we are experimenting with assigning weights to notes reflecting their contribution to the 'Gestalt' of a melody. Another approach could be to reduce derive harmonic progressions from melodies and apply similarity measures to these. Finally, we assume notation query input now, but more intuitive approaches are feasible.

Interactive Multimodal Search in Video Archives

Marcel Worring (University of Amsterdam, NL)

Searching for information in a large video archive is a difficult and challenging task. Users want to access the video archive using high-level concepts, but usually only low-level indices are available. In our view accessing video archives requires a four step proces. The first step is the off-line indexing stage to provide access points based on high-level concepts. As video is composed of multiple modalities this step should rely as much as possible on combining information from all the relevant sources. The next stage is the interactive part consisting of filtering using the indices to limit the dataset to an active set, browsing to find relevant examples, based on advanced visualization mechanisms, and a final ranking step based on the examples. Results will be presented on the TRECVID benchmark consisting of 60 hours of broadcast news video.

Video Google: A Text Retrieval Approach to Object Matching in Videos

Andrew Zisserman (Oxford University, GB)

An approach to object and scene retrieval will be described, which searches for and localizes all the occurrences of a user outlined object in a video. The object is represented by a set of viewpoint invariant region descriptors so that recognition can proceed successfully despite changes in viewpoint, illumination and partial occlusion. The temporal continuity of the video within a shot is used to track

the regions in order to reject unstable regions and reduce the effects of noise in the descriptors.

The analogy with text retrieval is in the implementation where matches on descriptors are pre-computed (using vector quantization), and inverted file systems and document rankings are used. The result is that retrieval is immediate, returning a ranked list of key frames/shots in the manner of Google.

The method will be demonstrated on two full length feature films ('Run Lola Run' and 'Groundhog Day').

Joint work of: Zisserman, Andrew; Sivic, Josef