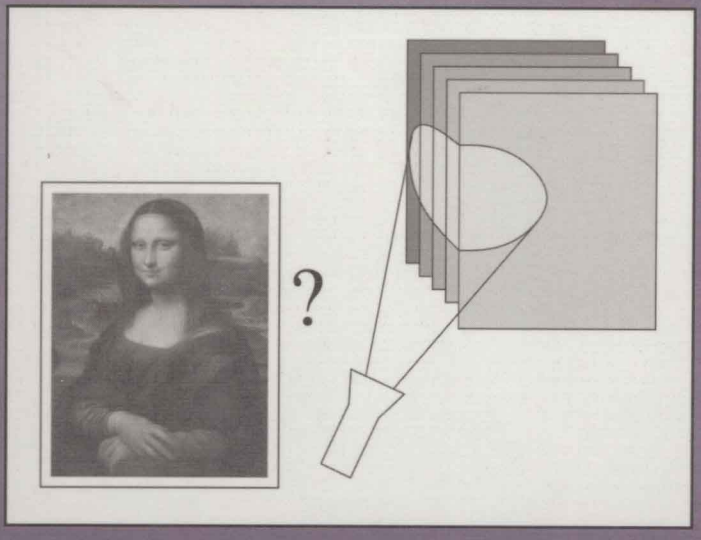
COMPUTATIONAL IMAGING AND VISION

State-of-the-Art in Content-Based Image and Video Retrieval

Edited by
Remco C. Veltkamp, Hans Burkhardt and
Hans-Peter Kriegel



Kluwer Academic Publishers

State-of-the-Art in Content-Based Image and Video Retrieval

Edited by

Remco C. Veltkamp

Utrecht University

Hans Burkhardt

University of Freiburg

and

Hans-Peter Kriegel

University of Munich

KLUWER ACADEMIC PUBLISHERS

DORDRECHT / BOSTON / LONDON

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 1-4020-0109-6

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands.

State-of-the-Art in Content-Based Image and Video Retrieval

Computational Imaging and Vision

Volume 22

Preface

Content-based image and video retrieval is concerned with retrieving images and video sequences on the basis of automatically derived features such as color, texture, and shape information that is present in the images. The need for efficient storage and retrieval of images is not new, but the increase in the number and size of digital image collections made the problems of image retrieval widely recognized. The shortcomings with traditional methods of indexing, often based on keywords, have led to the interest in retrieval on the basis of features that are automatically derived from the image content. The 1992 USA's NSF workshop on Visual Information Management Systems [1] was one of the events that gave this research area a push. The search for solutions is now an active area of research.

The emergence of content-based image and video retrieval has brought together and broadened the extend of a number of disciplines, such as image understanding, database indexing, object recognition, visual data modeling, feature extraction, visual querying, perception and cognition, and information visualization.

These developments motivated us to organize the Dagstuhl Seminar on Content-Based Image and Video Retrieval, 5-10 December 1999, Schloss Dagstuhl, Wadern, Germany [2]. The purpose of this seminar 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.

The past decade has witnessed the development of the first few commercial systems. These are still very limited in their functionality. The user requirements on image features can vary considerably, but they can be classified into three levels of abstraction: primitive features such as color or textures, logical features such as the identity of objects, and abstract features such as meaning or significance of the image [3]. While most current systems operate on the lowest of these levels, the user demands higher levels.

Therefore, the emphasis of the following seminar [4] 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.

Structure

This book is a result of the 1999 Dagstuhl Seminar on Content-based Image and Video Retrieval [2]. It contains a collection of works that represent the latest thinking in content-based image and video retrieval and cover a wide spectrum of areas. The chapters are a selection from the presentations at the seminar, and written for this book.

The chapters are grouped as follows:

- The first five chapters are dealing with features in image retrieval. To start off, chapter 1 discusses perceptual relevant features. Chapter 2 derives local image features from affine invariant regions. Chapter 3 is about local features integrated over rigid motions, and the evaluation with a Monte Carlo method. Chapter 4 treats 2D object retrieval by establishing correspondence between parts. Chapter 5 concludes this quintet with an overview of features used in 44 contemporary image retrieval systems.

- The next four chapters exploit, in different ways, probabilistic methods. Chapter 6 takes a probabilistic approach to classification and pose parameter estimation. Chapter 7, after an empirical evaluation, derives a framework that allows heterogeneous image models based on feature distributions. Chapter 8 is based on non-parametric density estimation of image feature clustering. Chapter 9 presents a logic-based retrieval system, where semantic information is transformed into a probabilistic object-oriented logic.

- The following four chapters are (in part) concerned with video retrieval. Chapter 10 introduces a multi-scale approach to image features, and a hierarchical motion classification for video. Chapter 11 describes, besides facial features, motion features derived from optic flow. Chapter 12 is about the hierarchical clustering of video key frames. In chapter 13, semantic information is derived from lower level features and signs.

- The last two chapters focus on indexing. Chapter 14 is treating similarity search that is adaptable to application specific requirements and user preferences. Finally, chapter 15 is about parallel nearest neighbor searching on a network of workstations.

Because this grouping is not strict and unambiguous, we have not made an explicit division of the book into parts.

Acknowledgment

We wish to thank all authors for their contributions, and their collaborative effort in producing an up-to-date state-of-the-art book in such a dynamic field. We thank also all other participants in the seminar for their presentations and lively discussions.

Last, but not least indeed, we thank the Dagstuhl organization for offering such a wonderful facility, and the Dagstuhl office for their perfect support.

References

[1] Ramesh Jain (ed.). NSF Workshop on Visual Information Management Systems, Redwood, CA, 24-25 Feb. 1992. In *Storage and Retrieval for Image and Video Databases, Proceedings SPIE 1908*, pages 198–218, 1993. Also appeared in SIGMOD Record 22(3), 57-75, 1993.

[2] Content-Based Image and Video Retrieval, Dagstuhl Seminar 99491, Schloss Dagstuhl, Wadern, Germany, 5-10 December, 1999. `http://www.dagstuhl.de/DATA/Reports/99491`.

[3] John P. Eakins and Margaret E. Graham. Content-Based Image Retrieval, A Report to the JISC Technology Application Programme. Technical report, Institute for Image Data Research, University of Northumbria at Newcastle, UK, January 1999. `http://www.unn.ac.uk/iidr/report.html`.

[4] Content-Based Image and Video Retrieval, Dagstuhl Seminar 02021, Schloss Dagstuhl, Wadern, Germany, 6-11 January, 2002. `http://www.dagstuhl.de/DATA/Reports/02021`.

Contents

Chapter 1

IMAGE CONTENT ANALYSIS AND DESCRIPTION

Xenophon Zabulis, Stelios C. Orphanoudakis
Institute of Computer Science, Foundation for Research and Technology - Hellas,
Vassilika Vouton, P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
and
Department of Computer Science, University of Crete,
P.O. Box 1470, GR-71409 Heraklion, Crete, Greece
{zabulis,orphanou}@ics.forth.gr

Abstract In this chapter the task of representing, describing, and analyzing visual information is discussed, in the context of image retrieval by content. Initially some basic specifications of the problem are presented and a classification of visual features, in a way compatible with human visual perception, is proposed. Through this discussion, it is realized that scale is an important attribute of visual content and a central issue in its description. Thus, a significant part of this chapter is devoted to the estimation and representation of primitive image content at different scales and a generic framework for this purpose is introduced. Finally, this chapter briefly considers the problem of how to derive and match descriptions of the visual content of an image in a perceptually correct manner.

Keywords: Image retrieval by content, visual information retrieval, scale-space, perceptual grouping, feature extraction, similarity matching.

1.1 Introduction

The large volume and variety of digital images currently acquired and used in different application domains has given rise to the requirement for intelligent image management and retrieval techniques. In particular, there is an increasing need for the development of automated image content analysis and description techniques in order to retrieve images efficiently from large collections, based on their visual content. Large collections of images can be found in many application domains such as journalism, advertising, entertainment,

R.C. Veltkamp et al. (eds.), State-of-the-Art in Content-Based Image and Video Retrieval, 1–19.

weather forecasting, map production, remote sensing, computer aided design, architecture, vision-based robot navigation, medicine, etc. Thus, an important functionality of next generation image database and multimedia information systems will undoubtedly be the search and retrieval of images based on visual content. In the near future, this functionality will also be supported by "intelligent" search engines used to access multimedia documents on the world-wide web.

Before a general solution to the problem of image browsing based on visual content can be found, there are many difficulties to be overcome. These difficulties stem primarily from the following facts or observations: 1) what constitutes image content in general is not well defined, 2) the degree of image similarity or dissimilarity is often context and goal dependent, 3) the types of images used and the requirements for content-based retrieval of such images are different for different application domains, and 4) mechanisms for selecting the image features to be used in content description and matching techniques are not well understood. Specifically, depending on the user's goal associated with a specific similarity search, a query by image content may be constructed based on either abstract or specialized image features. Image features used may also be global or local. The features used affect the precision of the response to a query by image content and the cardinality of the returned set of similar images. Precision and cardinality are also dependent on whether queries, using spatial and visual feature predicates, are exact or approximate. Exact queries require that a specific set of content descriptive criteria are necessarily satisfied, while approximate queries typically retrieve image with respect to their similarity with one or more visual examples.

The image type and context of use often determine those regions of interest and features that are characteristic of image content. The same visual stimuli may have different interpretations when observed in different contexts or by different observers. Furthermore, there is a semantic gap between the pictorial properties of an image and its linguistic interpretation. Thus, given these difficulties, the efficient, objective, and qualitative description of image content for the purpose of image similarity search is a complex task. A fundamental component of image content is structure, which resides at different scales between the scale defined by the sampling interval (pixel size) and the one corresponding to the size of the image itself. Therefore, in order to focus attention at structures of different sizes, it is important to have the ability to select the appropriate scale. If the size of a particular structure is known, the problem of estimating properties of this structure is simpler to solve. In general, scale selection is applicable to almost all image processing, feature detection, and image description tasks. Since scale selection appears to be an important factor in image content analysis and description, a significant part of this chapter

is devoted to the estimation and representation of primitive image content at different scales.

In describing the visual content of images and using such descriptions to retrieve similar images, the use of primitive image features may not be sufficient. One may also need to rely on more complex features obtained through perceptual grouping of primitive ones. In fact, a better understanding of human visual perception will undoubtedly contribute to the development of biologically relevant image content descriptions and more efficient mechanisms of image retrieval based on visual similarity. In this context, it is particularly important to define image similarity metrics, which correspond to known mechanisms of human visual perception. In this chapter, we also examine the role of human visual perception, and perceptual grouping in particular, in deriving descriptions of image content at a higher level than that afforded by primitive structure alone. Finally, this chapter briefly considers the problem of how to derive descriptions of the visual content of an image, which preserve information about its primitive, global, and perceptual features, while permitting salient regions within the image and selected features of such regions to carry an additional weight in image comparisons.

1.2 Problem Definition

The task of image retrieval by content may be subject to a number of requirements with regard to query types supported, retrieval precision, and the number of images retrieved. In all cases, one must first analyze and describe the visual content of the query image and match it to similar descriptions of images in a database. Before the central problem of image content analysis and description is addressed, a number of related problems and constraints are discussed below:

- *Image segmentation.* Segmenting an image into parts that are meaningful with respect to a particular application is critical in image understanding. However, segmenting an image into regions that correspond to distinct physical objects, using solely two-dimensional visual information is difficult or impossible to achieve. This is due primarily to the lack of three-dimensional models for every possible identifiable physical object and missing information regarding image acquisition parameters.

- *Motion and stereo vision* are sources of rich visual information. Visual cues provided by motion and stereo facilitate the extension of object boundaries, as well as the estimation of scene structure. On a semantic level, certain types of motion may constitute intense attractors, dominating an observer's attention. In static images there is no such visual cue. Similarly, stereoptic images can be used to estimate scene structure, thus contributing to the identification of distinct physical objects and scene understanding. This information is not available, in single images.

- *Lighting.* Knowledge of scene lighting plays an important role in the correct estimation of an object's reflectance spectrum. Human perception normalizes perceived spectra with respect to global scene illumination, a phenomenon known as "color constancy". However, in the general case of image acquisition, the scene illumination is neither known nor homogeneous. Specialized cases of color normalization, given certain assumptions about lighting conditions and / or an object's reflectance spectrum, exhibit interesting results, but the full reconstruction of an object's reflectance spectrum from a 3 band RGB image, is not trivial.

- *Object recognition.* The ability to identify specific objects in images would support the retrieval of semantically similar images. Images containing the same or "similar" objects, or even a contextually relevant object, may be considered as semantically related. Furthermore, object semantics may vary depending on the image observation goal and context, as well as the expectation of finding a particular object in a certain visual scene. For example, a tree trunk, which has been cut and is lying on the ground, may be characterized as a chair when taking a walk in the forest, while it could not be matched with any chair, stool or sofa model [1].

- *Context.* As already mentioned, the context of a query by image content and the type of images used have a strong effect on how the content of these images is described and compared. Contextual information and knowledge of the world are essential in deriving an appropriate image representation and may influence the role and significance of specific objects in such interpretations. Furthermore, the target class of images in a search and retrieval by visual content exercise may play an important role in determining which preprocessing methods are to be used for feature extraction.

- *Time.* Biological visual systems employ several physiological adaptation behaviors through time, such as lightness or chromatic adaptation [2], as well as motion adaptation [3]. Furthermore, given enough observation time, certain image features or details may be emphasized in the viewer's perception, depending on his / hers cognitive background and goal of observation. In this study a contextually uncommitted analysis of visual content is attempted, taking only into account only the early stages of visual perception.

- *Feedback.* Image feature extraction in biological vision systems may be adjusted depending on viewpoint, lighting conditions, query target, learning, adaptation and other factors. Feedback connections exist in the visual cortex, however their functionality has not yet been clearly understood. Certain image preprocessing methodologies may use feedback to improve feature extraction, but a generic framework for this is yet to be found.

In this chapter, a phenomenological thesis is adopted concerning the description of primitive image content and the evaluation of generic visual similarity. It is argued that visual content, once objectively represented, could also be appropriately interpreted, with respect to the context of use. However, most prominent of all problems mentioned above seems to be the quantification of qualitative visual attributes, such as image feature impression or the holistic perception of a scene.

1.3 Image Content Representation

Image features, including form and color or intensity distributions, as well as their spatial organization, compose primitive image content. However, some visual features reside in the perceptual domain, often defined by specific types of primitive feature arrangements. Some of these features may be detected by applying perceptual grouping rules [4] and are of strong descriptive power regarding the visual perception of a scene. Also, specific feature distributions may indicate regions of special interest in an image, such as regions attracting our preattentive attention, thus constituting qualitative information about image content. The scale at which features are observed is a central issue in image description. In this section, the importance of size in all cases of feature extraction is considered and some tools for dealing with feature scale are introduced. Methods for the estimation and classification of visual image features are also presented and discussed. In certain cases, an analogy is drawn between the applied methods and corresponding human visual perception mechanisms.

After a brief discussion of global image features and their role in image content description, this section emphasizes the estimation of primitive image features at selected scales and the representation of primitive feature distributions. Additional topics presented in this section are the perceptual grouping of primitive features into more complex ones and the role of regions of interest in images as attractors of attention in image retrieval by content.

1.3.1 Global features

A global statistical description of image features has been widely used in image analysis for image description, indexing, and retrieval. Such global feature descriptors include the image's color histogram, edge statistical infor-

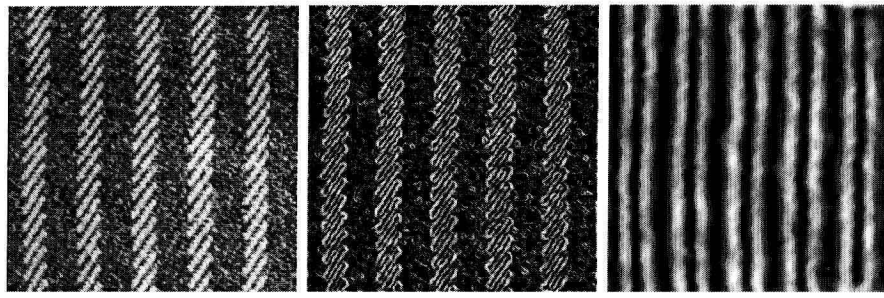


Figure 1.1 An original image (left) analyzed for edges at a fine (middle) and a coarse scale (right).

mation, wavelet or Fourier decomposition statistics etc.. Although these global attributes may be computed efficiently and often do succeed in capturing partial information about image content, they do not capture information about internal structure and cannot make use of any prior knowledge about a user's notion of image similarity based on specific interest in certain aspects of image content.

Despite the small discriminating power of such descriptors and the fact that they do not capture the spatial distribution of image features, their importance cannot be underestimated. Context based heuristics may be also used i.e. the detection of images containing man made structures may be achieved by searching for ones that are rich in straight line segments. In general, global image descriptors may offer important hints about overall visual appearance of an image, the image type, and certain possibly characteristic image properties. Using this information, images may be classified into categories, thus restricting the search space of image queries. Furthermore, knowledge of the image type often permits the selection of more suitable content analysis methods.

1.3.2 Primitive Features

Primitive features, such as edges, corners, blobs etc, model specific types of pixel distributions and constitute the "building blocks" of image content. They are highly correlated with the scale of observation and it is expected that their classification with respect to scale will contribute to the refinement of visual query formulation. As illustrated in Fig. 1.1, different aspects of visual content are observed as scale increases. Therefore, a full description of image content cannot be obtained by considering a single scale.

Observing the image input signal at all scales [5], using gradually coarser sampling, reveals the image content at each scale. Features at each scale can be detected by applying the appropriate operator at these scales. The operator response indicates the intensity (or probability) of feature presence at each pixel

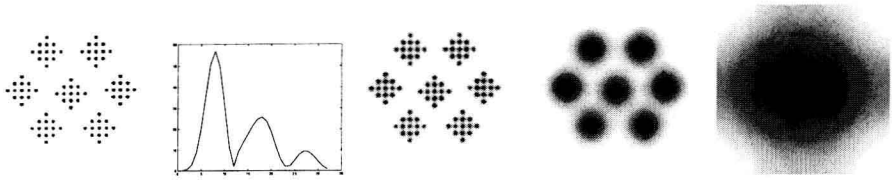


Figure 1.2 An image (left) and the plot of the blob scale selector response for the image point corresponding to the central dot. The three rightmost images present the image scales corresponding to the response function modes.

and with its use, feature points at each scale may be extracted. It may be scale normalized as in Eq. (1.1), where $\tau = \log t$ is the logarithmic scale parameter, $\vec{x}$ the pixel coordinates, $\mathcal{F}(\vec{x}, \tau)$ the feature operator and h is any strictly increasing function, chosen according to the nature of the feature detector.

$$\frac{1}{\int_0^\infty h(t\mathcal{F}(\vec{x}, \tau))\, d\tau} h\left(t\mathcal{F}(\vec{x}, \tau)\right). \tag{1.1}$$

In the example shown in Fig. 1.1, the feature detectors used, were the squared norm of the image gradient for edge detection, and the Harris [6] operator for corner detection. The example shown in Fig. 1.2, illustrates the response of the scale normalized Laplacian blob detector ($\mathcal{F}(\vec{x}, \tau) = |\frac{\partial^2}{\partial x^2} L(\vec{x}, \tau) + \frac{\partial^2}{\partial y^2} L(\vec{x}, \tau)|$, where L is the image scale-space), over all scales for the central point of an image, where the horizontal axis maps the logarithmic scale parameter τ. The bottom row presents the image scales corresponding to each of the detector's response modes. In general, the scale-normalized feature operator response reveals peaks, where the feature presence is most intense. Often more than one peak exists in a pixel's response over scale.

The intrinsic importance of scale in visual perception is observed in primate visual systems, where the sampled signal is passed as input to M and P retinal ganglion cells, that respond to spatial and temporal illumination change [7]. Different spatial "samplings" of ganglion cells are separately projected to the Lateral Geniculate Nucleus (LGN), and terminate in different regions of the visual cortex [8]. Coarse samples provided by the M ganglion cells project to magno cells in LGN, which are color-blind, high-contrast sensitive and with a fast neural response. In contrast, fine samples projected from P cells to parvo cells in LGN are color and low contrast sensitive, but have a slower response. Fig. 1.3 illustrates the described circuitry, representing M and P ganglion cells using black and white circles respectively. It is argued that the multiscale feature representation described in this section, is analogous to the primitive content representation observed in LGN.