International Workshop on Multimedia Information Systems and
Hypermedia, Tokyo, Japan, March 23-24, 1995.
M. Sakauchi and R. Jain, Eds.

Computer Vision and Multimedia Information Systems

Thierry Pun and Ruggero Milanese
Department of Computer Science, University of Geneva,
1211 Genève 4, Switzerland
pun@cui.unige.ch; http://cuiwww.unige.ch

Abstract

Computer vision offers a number of techniques that can be used in the context of pictorial information systems. To motivate this point of view, we first present the basic issues involved in computer vision and in multimedia information systems, and emphasize the differences between these domains. We then outline the contributions that computer vision can make to the development of efficient multimedia information systems, especially for handling queries by visual example (QVE).

In the second part of the paper, we concentrate on the matters of archiving and retrieving images from large databases, in the case of QVE. An approach is proposed that relies on two concepts, namely relevance and focus-of-attention. These mechanisms allow to locate and select the most pertinent information from images. They can be used at the image archival stage, to automatically determine which parts of the image should provide meaningful indices to be compiled into the image access tables. Relevance and focus-of-attention are also important at the retrieval stage, to efficiently select from a database the images that best match a given pictorial query.

1. Computer Vision for Multimedia Information Systems

1.1 Computer Vision

Computer vision (CV) aims at analyzing image data to understand the structure and content of scenes (e.g. [25] [28] [33] [38]). Images are organized along two or three dimensions, the third dimension being depth (e.g. [30]), or time as in the case of video sequences (e.g. [19]). The basic paradigm in model-based computer-vision is the interpretation of observations by comparison and matching with known models of the objects to be recognized.

Humans and animals are extremely skilled at using visual cues for inferring information about their environment. This fact has inspired many computer vision researchers, not only in the attempt to mimic the human visual system, but also by offering concrete pieces of information regarding the possible organization of a computer vision system. Hints from neurophysiology and cognitive psychology abound (e.g. [14] [36] [38] [57] [58]). There are for example experimental observations, such as the fact that object recognition is achieved in a few 100ms., leading to conjecture that less than a few tens neural cycles suffice to perform identification ([48]). It is also speculated that the human "model-base" might contain from 30'000 to 100'000 basic models ([6]). Despite this high number, humans can recognize objects under almost any affine or projective transformation. From a more general perspective, neurophysiology and psychology have revealed that the human visual system is structured into separate pathways, linking hierarchically organized modules that perform specific functions. The role of massive feedbacks is now acknowledged as, of course, the importance of parallel processing.

It is often recognized that despite initial hopes, computer vision has only achieved a limited success. Although there are now numerous applications of image analysis and vision techniques in research, industrial and daily life environments, the ultimate objective of realizing a general purpose computer vision system is still very far away. The goal of reproducing most of the human visual abilities thus remains elusive.

Computer vision is hard for many reasons: the problem is mathematically ill-defined, computationally intractable, and, last but not least, information extracted from real data is highly corrupted by noise ([25] [56]). Finding a general mathematical formulation of the vision problem still appears to be out of reach. The major current issue is to confront complexity by reducing input information as much as possible, and by using top-down constraints and feedback loops as early as feasible in the recognition process. Another trend consists of the active selection of input data, for example using moving cam-

era heads (e.g. [1]). Finally, in order to be able to handle real and poor quality data, it now becomes clear that domains of application have to be well specified, and heuristics used as often as possible.

1.2 Multimedia Information Systems

Multimedia information systems (MIS) aim at archiving and retrieving data from multiple sources of different types: text, music, speech, images (drawings, graphics, photographs), video sequences (news, movies) (e.g. [22] [31] [50] [60]). MIS research is multidisciplinary, as it involves database management, signal processing and computer vision, text and natural language processing, networking, human-media interaction.

The major operations that occur when dealing with a MIS are: the construction and precompilation of the database in order to optimize future searches, search operations to answer specific queries, browsing and selection amongst answers, and possibly search refinement. MIS are particularly well suited to exploratory "data mining", that is the search for relevant pieces of information in large knowledge repositories. The role of the human user in such operations is fundamental, at various stages: for selecting the appropriate data to be archived, for formulating and refining queries, and for browsing amongst responses to these queries.

1.3 Image and/or Spatial Databases

Images, be they static or dynamic video sequences, are certainly at the heart of any MIS. There is currently a distinction between spatial databases (e.g. [27] [37]) and image databases (e.g. [26] [29]), the former dealing more with the topological structure of the images, and the latter with their pictorial or semantic content. As techniques converge, this distinction will certainly become deprived of practical significance.

A central problem in information retrieval from pictorial databases is the handling of queries (e.g. [4] [13] [32] [50]). Structured, symbolic queries *à la* SQL are well suited to MIS where text descriptions of images exist, for example in the case of paintings. Such queries seem harder to use in the case of images with unrestricted content. In such situations, it is known that purely structural approaches to object recognition perform satisfactorily only in very restricted situations, such as for the analysis of line drawings, characters, or geographical maps (e.g. [49]).

The intuitive nature of data exploration is better suited to fuzzy than to symbolic queries. Such queries can be divided into queries by subjective descriptions (QSD) and queries by visual examples (QVE) ([35]). With QSD's, the user inputs a rough description, not necessarily pictorial, of the desired information. The query can for example consist of a series of adjectives describing certain subjective global attributes of images (such as "rather geometrical", "vividly colored", etc. [35]). In the case of QVE's, both objective and subjective *pictorial* criteria are provided to the system. Objective criteria are for example the dominant hue of the image to be retrieved, or the average size of the objects. More subjective criteria may for example be the dominant orientation of the patterns, or a hand-drawn sketch of a particular shape (e.g. [41]).

More and more examples of pictorial databases are appearing, commercially or within the research community. Regarding static images, applications can be found in the following domains: medical (Picture Archiving and Communication Systems - PACS, e.g. [46]), geographical (Geographical Information Systems - GIS [34] [49], earth-space observation [18]), museums and libraries (e.g. [5] [12] [35] [43] [47]), security (fingerprints, faces [3] [42]), artifact catalogs, home computing (handling of photographs), news agencies, etc.

Regarding video sequences, applications now revolve around automated channel selection, surveillance, or movies on demand. It will for example be possible to automatically browse through all channels offering a soccer match or a sumo wrestling contest ([15] [51] [61]). Such applications will certainly change focus with time, as TV channels will more and more be accompanied by data channels communicating the nature of the information being broadcasted. It will then not be anymore necessary to use complex techniques to locate channels offering a specific content, but rather to be able to eliminate unwanted information (such as commercial ads!).

MIS handling pictorial information have to deal with a number of data types (e.g. [22]). Raw binary data can be pixels, voxels, Binary Large Objects - BLOBS, video segments. In order to analyze such data, structures for representing image primitives and attributes of varying degrees of complexity are required. In addition, it is necessary to represent more global pieces of information, typically geometric, contextual or temporal (e.g. [21]).

1.4 Contribution of Computer Vision to Multimedia Information Systems

Several similarities exist between CV for objects location and recognition in images, and CV for MIS. Obviously, most basic techniques are the same. Also, in both cases, multiple data types have to be handled, and spatial queries to be processed.

There are however a number of differences between

CV for object recognition and for MIS. The major one is certainly the importance of human interaction in CV for MIS. The user is active, and is able to reformulate queries according to the results. In addition, and very importantly, there is no need to provide a perfect answer to a given search; several choices can be given, amongst which the user will browse. Another difference is that an image archive is not an object model-base, as could be obtained from a CAD/CAM design system. A complete, detailed representation is in general not available for describing the images in the archive. Finally, in MIS, it is conceivable that information be distributed into several repositories, whereas in object recognition there is usually only one modelbase.

What is then the role that computer vision techniques can play in MIS? We see significantly new contributions of CV to MIS in the case of queries by visual examples (QVE), and possibly in the case of queries by subjective description (QSD).

First, at the *archival* stage, the database has to be precompiled in order to optimize further searches. This involves extracting the most relevant primitives from the images, and organizing them ([11]). An additional problem is the automatic selection of keywords in order to build textual indexes ([24] [59]). Second, concerning the *human-media interactions*, due to the highly interactive and human-controlled nature of MIS, computer vision techniques might be necessary to translate queries by visual examples in terms of retrievable image primitives and attributes. Finally, at the *retrieval* stage, vision techniques have to be used for answering pictorial queries. In summary, only computer vision techniques can provide methodologies that allow search by content in pictorial MIS. The problem is a complex one, not only due to the need of interpreting fuzzy search criteria, but also since quite dissimilar data and models must be matched, and because of the efficiency constraint ([17]). Only a large use of domain heuristics will allow to satisfy all these requirements.

Interface design is usually a neglected issue in the development of a CV system for object recognition. In the context of MIS, care has to be taken to build multi-modal, intuitive interfaces, as users will certainly not be very computer-literate. Interesting developments for example revolve around automatic gesture recognition for interfacing with a computer system ([55]), in the use of virtual environments ([16]), or in speech recognition and natural language processing.

Amongst the various pictorial MIS existing or under development, most make only a parsimonious use of image analysis or computer vision techniques. For example, in medical PACS or museum applications, queries are generally answered on the basis of text records accompanying image data. This process requires manual image or video annotation, which may be tedious, slow, or too expensive. In order to improve the current systems, automatic generation of textual or symbolic description is thus necessary. This will certainly be a key factor for the development of computer vision techniques for multimedia applications.

2. Relevance- and Attention-based CV for Pictorial MIS

2.1 Handling Queries by Visual Examples

Queries by visual examples can be divided into two categories, *global* and *structural*. Queries in the first category are typically based on color or texture. For specifying a global color query, the user selects one or more dominant colors in Hue-Saturation-Lightness (HLS) space, as well as the relative percentage of each color that the retrieved images should contain (e.g. [35] [41]). For textural queries, the user may specify dominant orientations, or measures of activity of the textures to be retrieved (e.g. [4] [42]). Various traditional image analysis methods are available for handling such global queries. Current efforts aim at developing ergonomic user interfaces, as well as optimizing the search procedure ([17]).

More difficult is the matter of handling structural queries, where the user provides (possibly by drawing) a rough sketch of the "shape" that must be found in the database (e.g. [3] [4] [13] [24] [35] [41] [42]). In this case global techniques cannot be applied, since each shape defines a relational structure that must be spatially localized in the image. For each component of the shape, a correspondence must be found with a primitive in the matching image, and the consistency of all the spatial relations between these primitives with respect to the structural query must be verified. This can be formulated as a graph matching problem, which is known to be NP-complete.

In order to make this query mode feasible in practical applications, this general-case complexity must be reduced through some heuristics. The possibility of interaction with the user (cf. §1.4 above) clearly relaxes the constraint of finding the perfect match, and allows for multiple partial responses, amongst which the user may browse. Still, three basic problems can be identified for handling structural queries:

a) extracting basic image primitives and ranking them according to a measure of "quality";

b) locating and describing the most "interesting"

structures in the image;

c) finding a correspondence between the structural query and the primitives extracted at compile time from the images in the data-base.

In the following, we propose a three-steps strategy for solving these problems. At each step, heuristics are introduced to greatly reduce the computational complexity of the whole matching process:

a) ranking of image primitives by a *relevance* measure. This allows to choose the "best" primitives on which to start the matching process, and to greatly prune the search space;

b) locating "attention regions" containing structures of interest in the image, by means of a *focus-of-attention* mechanism. This also leads to a great reduction of the search space, since it permits to isolate groups of primitives likely to belong to a single object;

c) using fast indexing techniques for recognition, such as hashing methods. This requires to factor out variability of object appearances due to changes in scale and rotation.

In the rest of this paper, a general framework is presented for the three steps described above.

2.2 Relevance

Recognizing objects from a set of image primitives is a search problem of exponential complexity in the general case ([25] [56]). A major challenge in computer vision is therefore to select information that is relevant for recognition (e.g. [1] [10]). Significant efforts currently aim at developing efficient visual indexing schemes as a coarse but rapid preliminary recognition step ([20] [54]). Such schemes rely on finding those few key, or *relevant*, features that will drastically reduce the complexity of the search. A first problem is to find these features in a vast pool of image primitives. An additional issue is the impossibility to perfectly segment the target object: primitives are distorted, broken or simply missing. Finally, it is difficult to segregate object primitives from background ones. In this subsection, the definition and measurement of relevance values ([7] [8] [9]) is presented.

Let an image (or video frame) under consideration be segmented into P classes of primitives, such as line segments, circular arcs, and regions. The number of different primitive types is denoted by $p = 1 \ldots P$, (here $P = 3$). Let a *simple token* τ_i^p be a particular primitive of type p. Let a *token map* M^p be the set of all tokens of type p extracted from the image, together with their attributes and spatial relationships.

Objects or items of interest in an image may be composed of groups of simple tokens; one such local arrangement of simple tokens is called a *complex token* $T_j = \{\tau_{i1}^{p1}, \ldots, \tau_{in}^{pn}\}$. A complex token is therefore a structural entity, described by its components (segments, and/or arcs, and/or regions) and having its own coordinate system. Complex tokens are used for qualitative matching of objects with models, which is an operation typical of queries by visual examples (cf. §2.4).

For research purposes, we have constructed an artificial database of which representative images are shown in Figure 1; they are 256×256 color pictures of multiple 2-D objects (or 3-D objects presenting a stable 2-D view), lying on complex, textured backgrounds (gift wrap-paper). This type of images represents a difficult testbed since the highly textured background produces a large amount of primitives. In addition, the patterns that compose the background interfere with the foreground objects, so that no classic segmentation procedure can provide reliable primitives representing these objects (e.g. Figure 2, top, for segmentation examples). This research database currently holds about 60 different shape models superposed on different backgrounds, yielding about 200 composite images. Each image contains ≥ 1 objects, each of which is presented at a different rotation, position, and scale factor.

Figure 1: Typical samples from the current database of 200 images (originals are in color).

The line segment extraction is performed using a standard algorithm. A filter is then applied to remove segments shorter than a certain threshold (Figure 2.a, top). Circular arcs are obtained from a least-squares fit to the chains provided by the Canny edge detection algorithm (Figure 2.b, top). The region segmentation algorithm is based on two separate region growing mechanisms, that operate on the RGB color input image as well as on the Hue and Saturation planes. The results of the two segmentations processes are then fused, keeping the largest regions when overlapping is detected. The final result consists of a single region map (Figure 2.c, top).

For each simple token τ_i^p of type $p = 1 \ldots 3$ (line segments, circular arcs, regions) extracted from the input

image, the relevance $\rho(\tau_i^p) \in [0, 1]$ is defined by [7]:

$$\rho(\tau_i^p) = r(\tau_i^p) \cdot s(\tau_i^p) , \quad (1)$$

where $r(\tau_i^p)$ and $s(\tau_i^p)$ are respectively the *reliability* and the *significance* measures of τ_i^p, detailed below. High reliability indicates that a token is a meaningful entity, unlikely to have been generated simply by segmentation artifacts. The significance value measures the uniqueness of a token in the image; it is maximum when the attributes of τ_i^p make it unique in its type. The reliability and significance measures are obtained by analyzing some attributes computed for each primitive; the attributes employed for reliability and for significance are detailed below.

The attributes $A_r^p(\tau_i^p)$ used to compute the reliability of a token τ_i^p depend on the token map M^p to which it belongs. For line segments ($p = 1$), the two attributes $A_r^1(\tau_i^1)$, $r = 1, 2$ are length and contrast. Regarding circular arcs, the four attributes A_r^2, $r = 1 \ldots 4$ are radius length, arc length, contrast, fit error. Finally, for regions, the reliability attributes A_r^3, $r = 1 \ldots 3$, are area, average contrast, and standard deviation of the color distribution.

The attributes $A_s^p(\tau_i^p)$ used to compute the significance of a token also depend on its type. For line segments, the two attributes $A_s^1(\tau_i^1)$, $s = 1, 2$ are length and orientation. Regarding circular arcs, the three attributes A_s^2, $s = 1 \ldots 3$ are radius length, arc length, turning angle. Finally, for regions, the two significance attributes A_s^3, $s = 1 \ldots 2$, are area, and average intensity.

The reliability of a given token τ_i^p is the normalized (over the whole token map M^p) sum of all its reliability attributes r defined for its primitive type p:

$$r(\tau_i^p) = \sum_r A_r^p(\tau_i^p) / \sum_r \sum_i A_r^p(\tau_i^p) . \quad (2)$$

The significance measure is obtained by computing the sum of squared differences of a token's attributes with those of the other tokens of the same type:

$$s(\tau_i^p) = \sum_s \sum_{j \neq i} (A_s^p(\tau_i^p) - A_s^p(\tau_j^p))^2 . \quad (3)$$

Results of the relevance computation are presented in Figure 2, bottom. This figure shows that relevance allows to assess the respective "importance" of tokens of a given type.

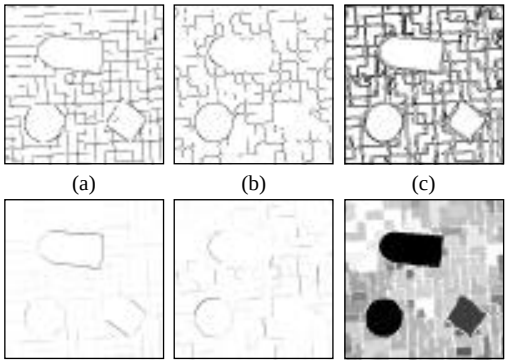


Figure 2: primitives and relevance measures $\rho(\tau_i^p)$. (Top) primitives extracted from the input image shown in Figure 1.a: (a) line segments; (b) arcs; (c) regions. (Bottom) representation of the relevance measures (darker pixels for tokens of higher relevance).

Relevances are computed independently for each type of primitive p. In order to obtain relevance values that may be compared across all primitive types, the initial relevances are statistically redistributed in [0,1] by separate histogram equalizations independently performed for each p. This yields *absolute relevance* values $\tilde{\rho}(\tau_i^p)$:

$$\rho(\tau_i^p) \rightarrow \tilde{\rho}(\tau_i^p) \in [0, 1] = Eq[\rho(\tau_i^p)] , \quad (4)$$

where equalizing functions $Eq[\bullet]$ are learnt for each type of token, over a set of similar images. A simple token of type p therefore has the same a-priori probability to be assigned a given relevance as any other token of type $p' \neq p$. After equalization, tokens of all P primitive types are ranked according to their $\tilde{\rho}$ value. Using this relevance evaluation, it is then possible to assess the relative "importance" of each image primitive for recognition, and in consequence to integrate primitives of various types.

2.3 Focus-of-attention

The visual attention module simulates the capability of biological visual systems to rapidly detect and locate "interesting" parts of a static retinal image, in order to reduce the amount of data for object recognition [39] [40]. Several criteria are used by the human visual system to evaluate the importance of a certain stimulus in the image. Some of them, described here, can be characterized as bottom-up, or data-driven. They are obtained by computing measures of saliency by comparing information extracted at each location with the rest of the image. Oth-

er criteria, rather top-down, involve some previously-stored knowledge. For instance, similarity of the stimulus with the shape of objects that are important for a certain task may be used, and/or their spatial relations with other objects (see [39] for extensions of the bottom-up algorithm to this type of *top-down* information).

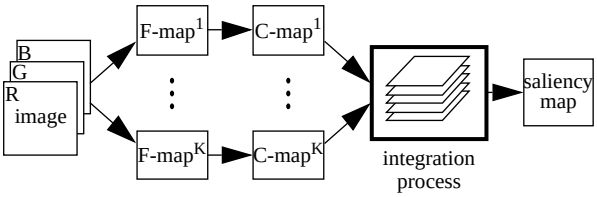


Figure 3: the bottom-up visual attention module.

The bottom-up subsystem is structured into three major stages, depicted in Figure 3. First, multiple retinotopic *feature maps* (*F-maps*) F^k, $k = 1 \ldots K$ are extracted from input images. The choice of these maps reflects some image properties that are computed in the visual cortex. Some of them represent chromatic information, and are obtained through color opponency filters *red-green* and *blue-yellow*. The other maps represent achromatic, high-frequency information, and are obtained through a bank of oriented, Gaussian 1st derivative filters. They encode information about the local edge orientation and magnitude, as well as local curvature.

The second stage of the attention system is represented by the extraction of the *conspicuity maps* (*C-maps*) C^k, one for each feature type k. The conspicuity maps represent K bottom-up measures of interest in the interval [0,1], at each location of the image. These measures are computed by convolving the feature maps with a bank of difference of oriented Gaussian filters, at multiple scales. The conspicuity map C^k is then obtained by computing the squared response at each location, and by taking the local maximum across different orientation and scales (see [39] for more details).

In the third stage of the system, the conspicuity maps are integrated into a single *saliency map*, defined as the average sum of the C-maps. However, a simple average sum directly computed from the original C-maps would average out all salient locations, rather then clearly detecting them. For this reason, an iterative non-linear relaxation algorithm is first applied to all C-maps. The updating rule is obtained by minimizing an energy measure, which has the effect of reducing noise, and enforcing regions that are active throughout multiple maps. At convergence, a binary mask is obtained by thresholding the saliency map in the middle of the range [0,1].

Figure 4 shows results obtained by the system on different types of input images. Even without any prior knowledge about objects of interest, the results successfully detect "irregularities" in the image, which correspond to objects that clearly stand out of a complex, textured background. In the context of pictorial MIS, this mechanism can be used to determine which are the "interesting" objects or parts of images that have to be used for archival or retrieval. Furthermore, in case of retrieval, it can be used to determine the most relevant components of the image query that must be found in the database.

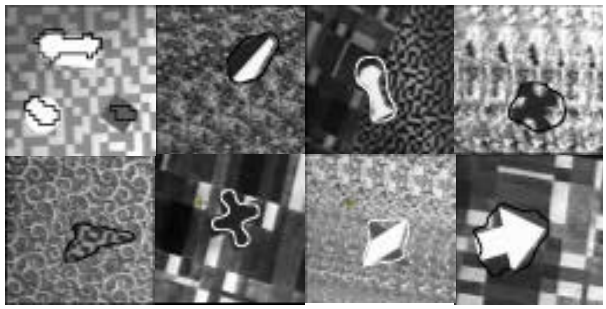


Figure 4: results of the attention system on some images of the database.

If $M > 1$ regions are detected by the attention system, multiple objects of interest are assumed to be present in the scene. In order to reduce the computational complexity of the following processing stages, the image can be split into M patches. In this way, only the information included in one patch is used for indexing at a given time. The results of the splitting procedure, implemented through a grass-fire algorithm, are shown in Figure 5.

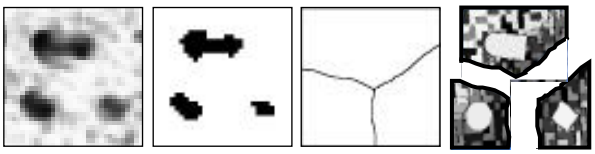
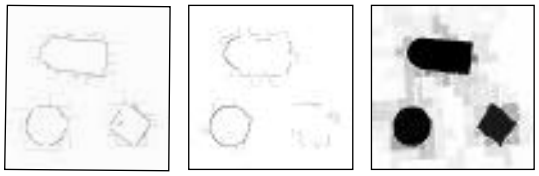


Figure 5: splitting the image according to the results of the attention mechanism, on the image shown in Figure 1.a; (a) result from the relaxation process; (b) attention regions; (c) objects' separation; (d) final patches.

The last stage for focalizing on the information necessary for accessing images or recognizing objects, consists of weighting the initial relevance values according to the location of the masks obtained by the focus of attention mechanism. A *proximity measure* $\pi(\tau_i) \in [0, 1]$ is computed for each primitive τ_i, with respect to the center of gravity of the attention region to which it belongs. $\pi(\tau_i)$ is maximal for close primitives, and decreases (exponentially) for more remote ones (cf.

Figure 6).



(a) (b) (c)

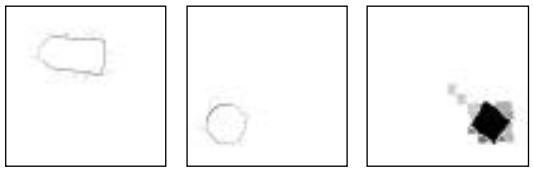
Figure 6: gating primitives from Figure 2 (top) with the proximity measure $\pi(\tau_i)$ of primitives to the center of each attention region (darker grey levels for shorter distance).

The relevance $\tilde{\rho}(\tau_i) \in [0, 1]$ of a single token τ_i is finally adjusted to take into account $\pi(\tau_i)$, yielding $\hat{\rho}(\tau_i) \in [0, 1]$:

$$\hat{\rho}(\tau_i) = (\tilde{\rho}(\tau_i) + \pi(\tau_i))/2 \quad . \tag{5}$$

Figure 7 shows the most characteristic segments and regions, i.e. those with the higher relevance $\hat{\rho}(\tau_i)$.

The extension of the present work to the handling of video sequences is described elsewhere ([23] [40]): using similar mechanisms, moving objects can be detected during an alerting phase, and tracked by means of a Kalman filter whose state vector describes positional features, such as a convex hull or a spline representation.



(a) (b) (c)

Figure 7: final relevance $\hat{\rho}(\tau_i)$ for the individual objects. (a) Line segments for the first object; (b) arcs for the second object; (c) regions for the third object (pixels darkness proportional to $\hat{\rho}$).

2.4 Matching and indexing

In QVE, indexing one or more pictures from the image database implies the ability to (rapidly) match a structural description provided by the user with the stored data. Various approaches exist for structural indexing, e.g. [2] [10] [20] [25] [30] [54]. In order to benefit from the relevance and focus-of-attention mechanisms, we have experimented with two different matching and indexing strategies, both operating with the complex token structures T_j introduced in §2.2. The strategy described below uses non-hierarchical complex tokens; in other words, each individual object is modeled as one complex token ([45]). Another strategy, described elsewhere [8], allows for each pattern to be described by a hierarchy of complex tokens.

In order to generate the hypothesis of a complex token T_j, the most relevant simple tokens τ_i extracted from the image (Figure 7) activate a local, purposive grouping process. This grouping process searches amongst the other highly relevant simple tokens those that would compose one of the stored complex tokens from the model base. Given an initial activating token that indexes an object model, the problem is to find other primitives that satisfy the geometrical relations included in the object model. To this end, the coordinate system of the activating primitive is first selected; the rotation and scale transformations specified by the relation parameters are then applied, leading to a formulation of the object model compatible with the activating primitive. The scene is finally searched for the missing primitives. This also allows to recover poorly segmented data, because expectations about missing tokens locally redirect a new segmentation with optimally defined parameters. Using this approach, a recognition rate of 100% was achieved on the training set of 60 shapes (objects without complex background), and 80% of the testing set (series of 200 composite images). Errors were due to incorrect regions from the focus-of-attention (5.7%), inaccurate segmentation (2%), incorrect recovery of the rotation angle (8.3%), and miscellaneous, such as objects too similar (4%). This matching and indexing approach is invariant to rotation, scale, translation, and is robust to disturbing background patterns or segmentation errors. Extensions to projective invariance are underway ([53]).

In order to use this approach in the context of QVE, all images stored in the database have to be processed as described in subsections 2.2 and 2.3. The most pertinent primitives are thus located, and their relevance quantified; these primitives constitute the model complex tokens on which query patterns will have to be matched.

3. Where do we go from here?

The integration of relevance and attentional mechanisms lead to a general framework that allows to fuse data from different sources, recover from poor segmentation, and handle uncertainty in an uniform manner. The relevance measure allows to detect the most pertinent primitives, quantifying their “importance” for recognition. The focus-of-attention spatially locates the most salient features in an image, and filters out irrelevant primitives.

In the context of MIS, the relevance and attentional

concepts can be used for images archival, images retrieval, and for human-media interaction. For images archival, these concepts allow to select important features, to rank them according to their pertinence, and to locate items of interest in the images. It is then possible to use these most important items as indexing keys for accessing pictures. Regarding images retrieval, relevance and attentional mechanisms appear as general paradigms for exploratory data mining, that provide a framework for qualitative indexing and matching which can be used for QVE and possibly for QSD. Finally, regarding human media interaction, relevance could be used in two ways. First, at the input of complex conjunctive queries, users could be asked to provide a relevance factor together with each component of their request. Second, for presenting results to the user, relevance could be used to rank the images retrieved from the database with respect to the query.

The results of the proposed technique have been described in the context of an artificial, still complex image database. Our current work consists of the integration of these concepts into an images archiving and retrieval system, applied to news photographs and to textile samples. We are also investigating learning techniques for automatic construction of object descriptions, aiming at precompiling appropriate indexing keys for efficient image retrieval.

To conclude, computer vision offers a number of techniques that can be used in the context of pictorial information systems. More specifically, queries by means of visual examples need to heavily rely on pattern recognition and object matching methods. Future challenges are in the domain of automated determination of indexing keys from large data sets, and in the development of new interaction models that integrate computer vision methods.

Acknowledgments. The authors wish to thank C. Rauber, Dr. J.-M. Bost, C. De Garrini, S. Startchik, for their contributions to this work. This research is partly sponsored by the Swiss National Research Foundation, the Swiss National Research Program "AI and Robotics", and the Swiss Priority Program in Informatics (grants 20-33467.92, 2000-040239.94, 4023-027036, and 5003-034278).

References

[1] Y. Aloimonos, Ed., Special Issue: Purposive, Qualitative and Active Vision, Comp. Vision, Graphics and Im. Proc.: Image Understanding, 56, 1, 1992.

[2] N. Ayache and O. Faugeras, "HYPER: a new approach for the recognition and positioning of two-dimensional objects", IEEE T-PAMI, 8, 1, Jan. 1986, 44-54.

[3] J. R. Bach, S. Paul, R. Jain, "A visual information management system for the interactive retrieval of faces", IEEE Trans. on Knowledge and Data Engineering, 5, 4, 1993, 619-628.

[4] R. Barber, W. Equitz, C. Faloutsos, M. Flickner, W. Niblack, D. Petkovic, P. Yanker, "Query by content for large on-line image collections", IBM Research Report, RJ 9408 (82660), June 29, 1993.

[5] H. Besser, "Imaging: fine arts", J. Amer. Soc. for Information Science, 42, 8, 1991, 589-596.

[6] I. Biederman, "Human image understanding: recent research and a theory", Comp. Vision, Graphics and Image Proc., 32, 1985, 29-73.

[7] J.-M. Bost, R. Milanese and T. Pun, "Temporal precedence in asynchronous visual indexing", Proc. 5th Int. Conf. on Computer Analysis of Images and Patterns (CAIP'93), Budapest, Hungary, Sept. 13-15, 1993, 468-475 (D. Chetverikov and W.G. Kropatsch, Eds., Springer-Verlag, Lecture Notes in C.S. 719, 1993).

[8] J.-M. Bost, "Active search for visual indexing in cluttered environments: from relevance to delays", Ph.D. Dissertation, No. 2656, University of Geneva, December 1993.

[9] P.-Y. Burgi and T. Pun, "Asynchronous image analysis: using the relationship luminance-to-latency to improve segmentation", J. Optical Soc. of America A (JOSA), 11, 6, June 1994, 1720-1726.

[10] A. Califano, R. Mohan, "Multidimensional indexing for recognizing visual shapes", IEEE Trans. PAMI, 16, 4, 1994, 373-392.

[11] T.P. Caudell, S.D.G. Smith, R. Escobedo, M. Anderson, "NIRS: Large scale ART-1 neural architectures for engineering design retrieval", Neural Networks, 7, 9, 1994, 1339-1350.

[12] A.E. Cawkell, "The British library's picture research projects", Advanced Imaging, Oct. 1993, 38-40.

[13] N.-S. Chang and K.-S. Fu, "Query-by-pictorial example", IEEE T-SE, 6, 6, Nov. 1980, 519-523.

[14] R. L. De Valois, K. K. De Valois, Spatial Vision, Oxford Science Publications, 1990.

[15] A. Del Bimbo, E. Vicario, D. Zingoni, "Supporting retrieval by contents of digital video sequences through spatio-temporal logic", Proc. 7th Int. Conf. on Image Analysis and Processing, S. Impedovo, Ed., Capitolo, Monopoli, Italy, Sept. 20-22, 1993. Published by World Scientific, S. Impedovo, Ed., 1994, 225-232.

[16] A. Del Bimbo, M. Campanai, P. Nesi, "A three-dimensional iconic environment for image database querying", IEEE T-SE, 19, 10, Oct. 1993, 997-1011.

[17] C. Faloutsos, M. Flickner, W. Niblack, D. Petkovic, W. Equitz, R. Barber, "Efficient and effective querying by image content", IBM Research Report, RJ 9453 (83074), August 3, 1993.

[18] U.M. Fayyad and P. Smyth, "Image database exploration: progress and challenges", 1993 AAAI Workshop on Knowledge discovery in databases, July 11-12, 1993, Washington DC, AAAI Press, Menlo Park, CA, Tech. Rep. WS 93-02.

[19] O. Faugeras, Three-dimensional Computer Vision: A Geometric Viewpoint, The MIT Press, 1993.

[20] P.J. Flynn, and A.K. Jain, "3D object recognition using invariant feature indexing of interpretation tables", Comp. Vision, Graphics and Im. Proc.: Image Understanding, 55, 2, 1992, 119-129.

[21] S. Gibbs, C. Breiteneder, D. Tsichritzis, "Data modeling of time-based media", ACM SIGMOD Record, 23, 2, June 1994, 91-102.

[22] S.J. Gibbs and D. Tsichritzis, Multimedia Programming: Objects, Environments and Frameworks, Addison-Wesley and ACM Press, 1994.

[23] S. Gil, R. Milanese, T. Pun, “Feature selection for object tracking in traffic scenes”, SPIE, Photonic Sensors and Controls for Commercial Applications - Intelligent Vehicle Highway Systems, Boston, USA, Oct. 31 - Nov. 4, 1994.

[24] Y. Gong, H. Zhang, H.C. Chuan, M. Sakauchi, “An image database system with content capturing and fast image indexing abilities”, Proc. Int. conf. on Multimedia Computing and Systems, May 14-19, 1994, Boston, MA, IEEE Comp. Soc. Press, 121-130.

[25] W.E.L. Grimson, Object Recognition by Computer, The MIT Press, 1990.

[26] W.I. Grosky, R. Mehrotra, “Guest editors' introduction”, IEEE Computer, Special Issue on Image Database Management, W.I. Grosky and R. Mehrotra, Eds., Dec. 1989, 7-8.

[27] R. H. Gueting, “An introduction to spatial database systems”, VLDB Journal, Special Issue on Spatial Databases Systems, H.-J. Schek, Ed., 3, 4, 1994, 357-399.

[28] R. M. Haralick, L. G. Shapiro, Computer and Robot Vision, Addision-Wesley, 2 Vols., 1993.

[29] S.S. Iyengar, R.L. Kashyap, “Guest Editors' Introduction: image databases”, IEEE T-SE, Special Section on Image Databases, 14, 5, May 1988, 608-611.

[30] A.K. Jain, P.J. Flynn, Three-Dimensional Object Recognition Systems, Series Adv. in Commun. Systems, Elsevier, 1993.

[31] R. Jain, “Visual databases and multimedia”, CVPR 1994 Tutorial, Seattle, WA, June 20, 1994. Accompanied by: W.I. Grosky, “Multimedia information systems: A tutorial”, CVPR 1994 Tutorial, Seattle, WA, June 20, 1994.

[32] T. Joseph, A.F. Cardenas, “PICQUERY: a high level query language for pictorial database management”, IEEE T-SE, Special Section on Image Databases, S.S. Iyengar and R.L. Kashyap Eds., 14, 5, May 1988, 630-638.

[33] R. Kasturi and R.C. Jain, Eds., Computer Vision: Principles, Advances and Applications, IEEE Computer Soc. Press, 2 Vols., 1991.

[34] R. Kasturi, R. Fernandez, M.L. Amlani, W.-C. Feng, “Map data processing in geographic information systems”, IEEE Computer, Special Issue on Image Database Management, W.I. Grosky and R. Mehrotra, Eds., Dec. 1989, 10-21.

[35] T. Kato, “Data-base architecture for context-based image retrieval”, SPIE Vol. 1662: Image storage and retrieval systems, 1992, 112-123.

[36] S.M. Kosslyn, O. Koenig, Wet Mind: The New Cognitive Neuroscience, Free Press, 1992.

[37] S.-Y. Lee, F.-J. Hsu, “Spatial knowledge representation for iconinc image database”, in: Handbook of pattern recognition and computer vision, C.H. Chen, L.F. Pau and P.S.P. Wang, Eds., World Scientific, 1993, 839-862.

[38] D. Marr, Vision, W.H. Freeman and Co., 1982.

[39] R. Milanese, H. Wechsler, S. Gil, J.-M. Bost, T. Pun, “Integration of bottom-up and top-down cues for visual attention using non-linear relaxation”, IEEE - CVPR 94 (Computer Vision and Pattern Recognition), Seattle, Washington, June 20-23, 1994, 781-785 (IEEE Computer Society Press).

[40] R. Milanese, S. Gil, T. Pun, “Attentive mechanisms for dynamic and static scene analysis”, accepted, Optical Engineering, July 1995.

[41] W. Niblack and M. Flickner, “Find me the pictures that look like this: IBM's image query project”, Advanced Imaging, April 1993, 32-35.

[42] A. Pentland, R.W. Picard, S. Sclaroff, “Photobook: tools for content based manipulation of image databases”, MIT Media Lab., July 25, 1994.

[43] D. Pountain, “The fine art of CD-Rom publishing”, Byte, June 1994 (about Microsoft’s Art Gallery).

[44] T. Pun, J.-M. Bost, R. Milanese, C. Rauber, S. Startchik, “Selecting relevant information and delaying irrelevant data for objects recognition”, AAAI Fall Symposium Series, Relevance Workshop, New Orleans, Louisiana, 4-6 Nov. 1994; AAAI Press, 168-172.

[45] T. Pun, C. Rauber, S. Startchik, R. Milanese, “Transforming an image into dataflows of relevant primitives for objects location, reconstruction and indexing”, Vision Interface 95, Quebec City, Canada, May 15-19, 1995.

[46] O. Ratib, Ed., Special Issue: Multimedia techniques in the medical environment, Computerized Medical Imaging and Graphics, 18, 2, 1994.

[47] R.P.C. Rodgers, S. Srinivasan, “On-line images from the history of medicine (OLI): Creating a large searchable image database for distribution via World-Wide Web ”, First Int. WWW Conference, Geneva, 25-27 May, 1994 (http://www.nlm.nih.gov:8002/paper/paper.html).

[48] A. Rosenfeld, “Recognizing unexpected objects: a proposed approach”, Int. J. of Pattern Rec. and Artificial Intell., 1, 1, 1987, 71-84.

[49] N. Roussopoulos, C. Faloutsos, T. Sellis, “An efficient pictorial database system for PSQL”, IEEE T-SE, Special Section on Image Databases, S.S. Iyengar and R.L. Kashyap Eds., 14, 5, May 1988, 630-638.

[50] M. Sakauchi, “Database vision and image retrieval”, IEEE Multimedia Journal, Report Section, Spring 1994, 79-81.

[51] T. Satou, M. Sakauchi, “Video acquisition on live hypermedia”, IEEE Multimedia Conf. 1995, May 1995, USA.

[52] F. Stein and G. Médioni, “Structural indexing: efficient 3D object recognition”, IEEE - Trans. PAMI, 14, 2, Febr. 1992, 125-145.

[53] S. Startchik, C. Rauber, T. Pun, “Recognition of planar objects over complex backgrounds using line invariants and relevance measures”, Europe-China Workshop on Geometrical Modelling and Invariants for Computer Vision, Xi'an, China, April 27-29, 1995.

[54] F. Stein, G. Médioni, “Structural indexing: efficient 3D object recognition”, IEEE Trans.-PAMI, 14, 2, 1992, 125-145.

[55] K. Takahashi, S. Seki, R. Oka, “Spotting recognition of human gestures from motion images”, Time Varying and moving objects recognition, 3, Proc. 4th, Int. Workshop, Florence, V. Cappellini, Ed., It., June 10-11, 1993, 65-72.

[56] J.K. Tsotsos, “Analyzing vision at the complexity level”, Behav. & Brain Sciences, 13, 1990, 423-469.

[57] R. Watt, Understanding Vision, Academic Press, 1991.

[58] H. Wechsler, Computational Vision, Academic Press, 1990.

[59] J. Yamane and M. Sakauchi, “A construction of a new image database system which realizes fully automated keyword extraction, IEICE Trans. Inf. and Systems (Japan), Vol. E76D, No. 10, Oct. 1993, 1216-1223.

[60] A. Yew-Hock, A. D. Narasimhalu, S. Al-Hawamdeh, “Image information retrieval systems”, in: Handbook of pattern recognition and computer vision, C.H. Chen, L.F. Pau and P.S.P. Wang, Eds., World Scientific, 1993, 719-740.

[61] H.-J. Zhang, S. Y. Tan, S.W. Smoliar, G. Yihong, “Automatic partitioning and indexing of news video”, Multimedia Systems, 2, 1995, 256-266.