INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

The Methodology and Practice of the Evaluation of Image Retrieval Systems and Segmentation Methods

Ian Jermyn — Cián Shaffrey — Nick Kingsbury

N° 4761

March 2003

THÈME 3

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--4761--FR+ENG

The Methodology and Practice of the Evaluation of Image Retrieval Systems and Segmentation Methods

Ian Jermyn , Cián Shaffrey* , Nick Kingsbury†

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projets Ariana

Rapport de recherche n° 4761 — March 2003 — 34 pages

Abstract: Content-Based Image Retrieval is important for two reasons. First, the oft-cited growth of image archives in many fields, and the rapid expansion of the Web, mean that successful image retrieval systems are fast becoming a necessity if the mass of accumulated data is to be useful. Second, database retrieval provides a framework within which the important questions of machine vision are brought into focus: successful retrieval is likely to require genuine image understanding. In view of these points, the evaluation of retrieval systems becomes a matter of priority. There is already a substantial literature evaluating specific systems, but little high-level discussion of the evaluation methodologies themselves seems to have taken place. In the first part of the report, we propose a framework within which such issues can be addressed, analyse possible evaluation methodologies, indicate where they are appropriate and where they are not, and critique query-by-example and evaluation methodologies related to it. In the second part of the report, we apply the results of this analysis to a particular dataset. The dataset is problematic but typical: no ground truth is available for its semantics. Considering retrieval based on image segmentations, we present a novel method for its evaluation. Unlike methods of evaluation that rely on the existence or creation of ground truth, the proposed evaluation procedure subjects human subjects to a *psychovisual* test comparing the results of different segmentation schemes. The test is designed to answer two questions: does consensus about a 'best' segmentation exist, and if it does, what do we learn about segmentation schemes for retrieval? The results confirm that human subjects are consistent in their judgements, thus allowing meaningful evaluation.

Key-words: evaluation, methodology, content-based, retrieval, semantics, image database, image segmentation, psychovisual test

* Signal Processing Laboratory, Department of Engineering, University of Cambridge, UK. (cws23@eng.cam.ac.uk)
† Signal Processing Laboratory, Department of Engineering, University of Cambridge, UK. (ngk@eng.cam.ac.uk)

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

La Méthodologie et la Pratique de l'Evaluation de Systèmes de Recherche en Bases de Données Image et Méthodes de Segmentation

Résumé : La recherche d'images par le contenu est importante pour deux raisons. Premièrement, la croissance d'archives d'images fréquemment citée dans beaucoup d'applications, et l'expansion rapide du Web, signifient qu'il est nécessaire d'utiliser des systèmes de recherche efficaces pour les bases de données afin que la masse de données accumulée soit utile. Deuxièmement, la recherche dans les bases de données image pose des questions importantes liées à la vision par ordinateur : une recherche efficace demande une véritable compréhension des images. Pour ces raisons, l'évaluation des systèmes de recherche dans les bases de données image devient une priorité. Il existe déjà une littérature importante évaluant des systèmes spécifiques, mais peu de discussions sont publiées sur les méthodes d'évaluation en soi. Dans la première partie de ce rapport, nous proposons un cadre dans lequel ces sujets peuvent être abordés, nous analysons des méthodologies d'évaluation possibles, indiquant quand elles sont pertinentes et quand elles ne le sont pas, et nous critiquons la technique "query-by-example" et les méthodes d'évaluation qui s'y rapportent. Dans la deuxième partie du rapport, nous appliquons les résultats de cette analyse à une collection spécifique d'images. Cette collection est problématique mais typique: il n'existe pas de vérité terrain sémantique. Considérant la recherche fondée sur la segmentation d'image, nous présentons une nouvelle méthode pour son évaluation. Contrairement aux méthodes d'évaluation qui reposent sur l'existence ou la création d'une vérité terrain, la méthodologie proposée utilise des sujets humains pour un test psychovisuel qui compare les résultats des différentes méthodes de segmentation. Le test est conçu pour répondre à deux questions : existe-t-il une segmentation "meilleure" que les autres et si oui qu'apprenons-nous des méthodes de segmentation pour la recherche dans des bases de données image? Les résultats confirment la cohérence des jugements humains, permettant ainsi une évaluation significative.

Mots-clés : évaluation, méthodologie, contenu, recherche, sémantique, base de données image, segmentation, test psychovisuel

Contents

1 Introduction

It is a commonplace that the growth of the Web and the ever-growing collections of digital images in many different fields renders pressing the need for genuinely content-based image retrieval systems. In addition to this practical necessity, the task of image retrieval provides a framework within which to view the important problems of machine vision. Successful content-based image retrieval will require genuine image understanding, and indeed retrieval can be viewed as pullback by the image understanding arrow. In consequence, the ability to evaluate image retrieval systems in a way that does not depend on the opinion of the evaluator becomes a matter of priority.

For collections of images for which there is no well-defined semantics, and hence nothing to which to compare the results of retrieval (a situation that occurs frequently), evaluation methodology becomes a particularly murky area, and little high-level discussion of methodological issues seems to have taken place. The aim of section 2 is to shed light on the evaluation of retrieval systems by proposing a framework within which the issues can be described. We perform an analysis of possible evaluation methodologies in scenarios with different degrees of structure, indicating where they are appropriate and where they are not, and critique query by example and the evaluation techniques associated with it.

In section 3, we apply the results of this analysis to a particular dataset, the Bridgeman Art Library collection described in section 1.1. Using this dataset, we implement the evaluation of retrieval systems that use segmentation algorithms to generate indices for retrieval. In the absence of a well-defined semantics for the dataset, we develop and perform *psychovisual* experiments to test the performance of the algorithms. These tests can equivalently be viewed as an evaluation of the segmentation algorithms themselves. Our results indicate that, despite the lack of a well-defined semantics for this dataset, a consensus does emerge about what constitutes a 'good' segmentation. The existence of this consensus is far from obvious *a priori*, and is in itself a significant result. In addition, we obtain a consistent ranking of the segmentation schemes evaluated. The way is thus open for a more detailed study of the factors underpinning this consensus and ranking.

The work described in this report was carried out within European project MOUMIR (Models of Unified Multimedia Information Retrieval). It is part of a study of evaluation for image retrieval systems that will include also the IGN dataset described in 1.1.1. The work on evaluation using the IGN dataset outlined in this report is underway, and will be reported at a later date.

1.1 Datasets

Retrieval systems cover an ever-expanding range of application areas. Some work with very limited 'scenes',[1] and hence with narrowly defined sets of images with precise semantics, while others work with generic scenes whose semantic content seems unbounded. We will bear the following two examples in mind as we proceed. These examples will serve to make the discussion concrete; they represent two extremes of database usage and evaluation.

[1] We use the word 'scene' to denote whatever is of interest in the image: this can range from the position and identity of objects in a real (or imagined) 3D scene, to abstract, precisely defined symbols or even camera parameters.

1.1.1 L'Institut Géographique National

This dataset is a collection of aerial images of the Ile-de-France region around Paris created by the Institut Géographique National (IGN), the French Mapping Institute. An example is shown in figure 1(a). Each of the images shows a land area of 4.6km by 4.6km, and is 500 by 500 pixels, so that the resolution is about 9m/pixel. There are many classes of statement that one could consider making about such images, but one of the most important consists of statements about land use. Such statements essentially involve a map from the image domain into a finite set of classes: 'forest', 'urban area', 'agricultural field' and so on. A set of such maps is available from the Institut d'Aménagement et d'Urbanisme de la Région d'Ile-de-France (IAURIF), who compiled them (independently of the IGN images) using field studies and existing cartography. The land use map for part of the image in figure 1(a) is shown in 1(b).

1.1.2 Bridgeman Art Library

This dataset is a collection of images of paintings from the Bridgeman Art Library (BAL). Based in the United Kingdom, BAL is a commercial art library supplying electronic and hard copy images to magazines, newspapers, designers and others. The images are realistic in intent, but in many cases the colours and forms do not correspond to 'photographic realism'. It is very hard to characterize the queries faced by the staff at BAL. The queries are often phrased at a very high semantic level, and the process of answering queries is complex, involving prolonged interaction with clients. Two example images are shown in figure 1(c) and (d).

Figure 1: (a): an example of the IGN aerial images (© IGN); (b): the land use map for part of this image (© IAURIF); (c) and (d): examples of the BAL images (© BAL).

2 Methodological Analysis

In this section, we perform an analysis of evaluation methodologies for retrieval systems. There is already a substantial literature evaluating various specific retrieval systems, but little analysis of the underlying methodology seems to have taken place. Much of the work in evaluation uses query by example and "relevance" classes of images (see, among many others, [13, 18, 2, 3, 1, 23, 15] and the many references in the reviews [21, 22, 19, 16]). One of the main arguments of this section is that both these techniques are flawed conceptually, and that used together they give the appearance of objectivity without the substance.

The rest of this section is structured as follows.

Section 2.1 We discuss general issues related to the evaluation of retrieval systems *in situ*, and discuss the assumptions necessary to evaluate them "in the laboratory". We describe a 'database schema', an abstraction of a retrieval system that contains all the elements necessary to discuss evaluation, and then introduce the notion of a 'knowledge scenario'. Different knowledge scenarios correspond to different states of knowledge regarding the elements of the database schema. We also describe and critique the assumptions implicit in query by example.

Section 2.2 We discuss the quantities necessary for evaluation, and introduce the idea of an 'evaluation context', the space in which an evaluation takes place. Starting from the database schema, we describe the three evaluation contexts available in the case of retrieval systems.

Section 2.3 We analyse which evaluation contexts make sense in which knowledge scenarios. We continue the discussion of query by example by analysing and critiquing the typical evaluation of query by example using "relevance" classes.

Section 2.4 We apply the analysis to the two datasets described in section 1.1, and describe concrete evaluation methods for these datasets. The analysis applied to the BAL dataset will be taken up in detail in section 3, where we implement the results of the analysis.

Section 2.5 We summarize the conclusions of this section.

2.1 Database Schema and Knowledge Scenarios

Database systems are intended for a particular situation. Particular individuals will access the database and try to retrieve images for a particular purpose. The database will either give the individuals what they want with ease and grace, or it will not. This implies that the fullest way to evaluate such systems is to study the performance of the system *in situ*, through the reactions of users, surveys, work rates, and so on, and to come to a conclusion from this data. Whether there is any consistency in the evaluation of different systems across different applications and datasets, or even across different work environments and personnel for the same application and dataset, is an empirical question, to be answered by experiments and not by assumption. If little consistency between the evaluations across different environments were to be found, there would be no well-defined environment-independent sense in which one system could be said to be better than another.

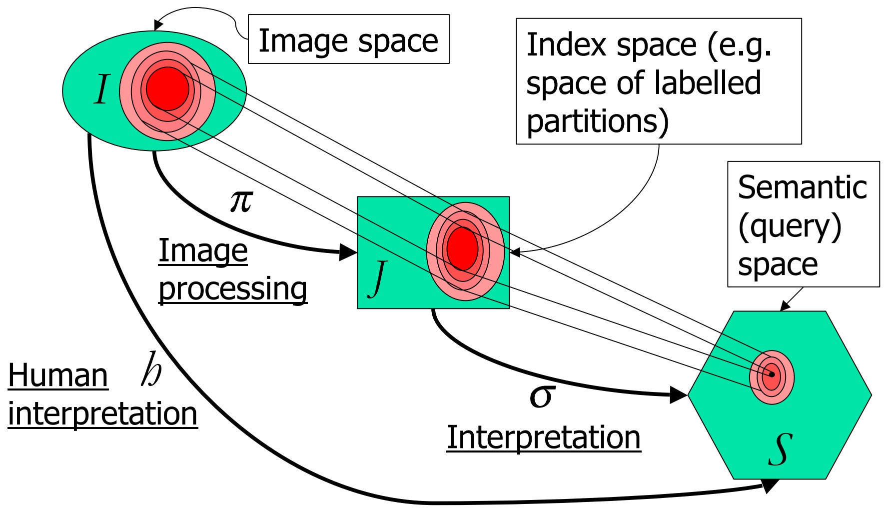


Figure 2: Database schema

This situation is not very satisfactory. It implies that we can say very little about the performance of image retrieval systems without conducting large and impractical experiments. If we wish to go further in the absence of such experiments, we must *assume* that we can abstract the idea of performance away from the environment in which the software will be used without seriously affecting the evaluations themselves. To make this abstraction, we must find some way to duplicate the evaluation of the 'average relevant user'. By doing so, we are effectively 'integrating out' the variables in which we have no interest, in this case the application environment and the diversity of users, leaving a marginalised performance measure.

We clarify these ideas using figure 2, which shows an abstract representation of a retrieval system. The figure shows a series of spaces and a number of arrows between them. The arrows can be interpreted either as maps between the spaces themselves, or as maps between the probability measure spaces on those spaces (i.e. conditional probabilities). We will introduce the components of this diagram one by one.

The 'image space', I, is the space of images. It contains a finite subset: the set of images in the database. (We will not distinguish carefully between I and this subset. This creates no confusion.) Subject to a query, the output of the retrieval system will be a subset of this subset.

The 'semantic space', S, can be thought of as the space of atomic statements we might like to make about the scene for a particular application, and will concern some part of the image formation process. Examples are statements about the real world scene of which a photograph was taken; the objects represented in a blueprint or other design document; the 3D scene represented in a painting;

the properties of the painting itself as an artefact; camera parameters used to acquire an image; and so on.

The map h from I to S represents the interpretation of the images by a user. In general, for each user, their interpretation will take the form of a map between the spaces of probability measures on I and S. In other words, given an image in I, the user's interpretation will map it to a probability distribution over S that represents the uncertainty of the interpretation of that user. To obtain the 'average relevant user', we should then combine the distributions from different users to obtain a distribution over the space of interpretations of many users, and then marginalise over the different users to obtain a distribution over interpretations. If the entropy of this marginalised conditional distribution is not much higher than the entropy of the distributions for each user individually (there is 'consensus'), then the notion of an 'average relevant user' is useful. Otherwise we are faced with performing the type of experiment discussed at the beginning of this section. If in addition to consensus, the resulting marginalised distribution is narrowly peaked about one value, then the map between probability spaces can usefully be replaced by a map between the spaces themselves. We will talk as though this is the case, although little is changed in the subsequent arguments by relaxing this restriction.

Once I, S and the arrow h are defined, we can talk about the ideal output of the retrieval system as follows. A query is assumed to specify a subset of S.[2] Given such a query, the map h can be used to pullback[3] the subset to I, thus specifying a set of retrieved images.[4]

The other space indicated in the figure is the 'index space', J. This, and its associated maps π and σ, constitute a factorization of the machine, as opposed to human, image understanding chain. The map π represents the image processing applied to the images in the database to generate indices for retrieval. Usually the high-dimensional image space is projected to a much lower-dimensional space that it is hoped nevertheless still captures the relevant information about the images. The latter phrase means that there exists a map τ such that $h = \tau\pi$. This cannot happen in particular, if two images map to the same point in J under π that are mapped to different points in S under h. In other words, the index space and the image processing map must be 'fine enough'.

The semantics map, σ, represents further processing that maps the index space to statements about the image. The ideal situation is that $\sigma\pi = h$. If this were the case, then machine retrieval would be the same as ideal human retrieval: pullback from a query using the arrow $(\sigma\pi)^* = \pi^*\sigma^* = h^*$. One may wonder why we need to factorize the map $\sigma\pi$ and create the space J. The answer is that construction of a complete map $\sigma\pi$ that approximates h is usually impossible at the present state

[2]The specification of this subset can be done in a variety of ways, text-based queries and example images being the most usual. Some of these methods may function better than others in different application environments, but given our goal of abstracting away from issues associated with these environments, we will suppose that the translation to S is perfect. Note that consequently we do not examine such techniques as 'relevance feedback', which can be viewed as methods for perfecting the query. This therefore excludes the discussion of such performance characteristics as how fast a user may reach their final query, and of what limits the feedback process may put on the trajectories possible in the semantic space.

[3]For an arrow $f : A \to B$, we use the notation $f^* : 2^B \to 2^A$ to indicate pullback: $\forall Y \subset B : f^*(Y) = \{a \in A : f(a) \in Y\}$., and the notation $f_* : 2^A \to 2^B$ to indicate push forward: $\forall X \subset A : f_*(X) = \{f(a) \in B : a \in X\}$.

[4]Many methods introduce a metric on S. We consider the use of a metric either as specifying a less restrictive query or as a probabilistic version of the current discussion. Consequently our arguments still apply.

of development. The space J thus represents as far as we can presently go along the line between images and semantics.

Problems arise in building and evaluating retrieval systems because one or the other of the above quantities is difficult, if not impossible to characterize.[5] These problems have a well-defined order, in the sense that the inability to solve a problem earlier in the list renders the subsequent problems moot. The various stages in this list will be called 'knowledge scenarios'.

2.1.1 Knowledge Scenario 1

In this scenario, we cannot characterize S explicitly. Clearly, in this case neither h nor σ are characterizable explicitly. It is usually only possible to construct S if a limited number of anti-atoms can be found in terms of which to express the statements as conjunctions. In the case of the IGN images, for example, conjunctions of statements of the form: "such-and-such region in the image domain corresponds to land use of such-and-such type", are enough to express all relevant queries. No such simple characterization exists for the images from the BAL dataset, and indeed the semantic space seems unbounded. The BAL dataset and the queries typically made of it are a good example of this first scenario.

2.1.2 Knowledge Scenario 2

In this scenario, we know how to characterize S, but we do not know how to characterize h explicitly. Difficulties with characterizing h fall into two broad classes. There may be consensus among users about the semantics in a probabilistic sense, while still being a great deal of uncertainty about the interpretations. In this case, h must be described using conditional probabilities. It may be difficult to obtain the information necessary to describe this distribution. In addition, I and S may be too large to allow explicit construction of the map, requiring prohibitive resources of money or time. In the IGN case, exhaustive enumeration of the semantics is possible. There is one land-use map giving the value of h for every image in the database; these were created by human operatives. It is however easy to imagine increasing the complexity of the semantics or the number of images in such a way that exhaustive enumeration would become impossible. We might then know the statements that we wish to make, but be unable to ascertain whether they are true or not of a given image.

2.1.3 Knowledge Scenario 3

In this scenario, we know how to characterize S and h, but we cannot construct a successful version of the arrow $\sigma\pi$. This scenario is a little different from the previous two. There, we could not characterize what we wanted to retrieve. Here, we can characterize it, but we cannot hope to duplicate it. Of course we can always construct maps that take us from J to S. The problem arises when none of the maps we can construct produces a $\sigma\pi$ remotely close to h. Our attempts at retrieval are then doomed to failure. An example is the following. Assume we have a set of images and that we wish to make statements of the form "In the scene that generated this image, a human being occupied

[5] "Characterize" here means 'describe algorithmically', either by enumeration or by some other procedure. We use the terms 'characterize', 'construct', and 'describe' interchangeably.

a volume that projected to region R in the image domain". The semantics is very clear, and we can construct h simply by inspection, or through prior knowledge of the scenes from which the test images were generated. It is however a difficult task to do this automatically for general classes of images.

2.1.4 Knowledge Scenario 4

In this scenario, we can characterize both S and h, and we can construct reasonably successful maps $\sigma\pi$. The IGN dataset, coupled with queries about land use, form a good example of this scenario. We have a well-defined semantic space, described in section 2.3.4, a well-defined arrow h, given by the images produced by IAURIF, and segmentation algorithms that do a reasonable job of partitioning the image domains into the correct subsets. Other examples that fall into this knowledge scenario include [4] and [24], who use multiple expert observers to agree on ground truth in the context of medical imagery, and [7] who use carefully created ground truth to test range-finding algorithms.

2.1.5 Query By Example

It is unfortunately all too often the case that we find ourselves in knowledge scenario 1. One common response to this is to attempt to circumvent the need for a semantics as follows. In the absence of any well-defined maps, one instead assumes that the user interprets images via an unknown map h to an unknown S, and that he has in mind an unknown query $q \subset S$. One then allows the user to select a small subset $i \subset I$ of images from the database that 'represent' the unknown query q, which means in principle that $i \subset h^*(q)$. Now that one has a set of images to work with rather than a query, one can apply the arrow $\pi^*\pi_*$ to generate a set of retrieved images in I. Thus query by example removes the need for the unknown quantities S and h, and hence σ, by employing the user himself to translate from the unknown query to a set of 'example' images.

Note that if there exists an arrow σ such that $h = \sigma\pi$, then

$$h = \sigma\pi \Rightarrow h^* = \pi^*\sigma^* \Rightarrow \sigma^* = \pi_* h^* \Rightarrow h^* = \pi^*\pi_* h^*. \tag{1}$$

It then follows that $i \subset \pi^*\pi_*(i) \subset h^*(q)$, so that if σ exists, the retrieval process will return more images with the same semantics. Thus the existence of σ such that $h = \sigma\pi$ means that π divides I into equivalence classes that are a refinement of those generated by the unknown h: images i and i' such that $\pi(i) = \pi(i')$ necessarily satisfy $h(i) = h(i')$.

Query by example raises a number of difficulties. Note that a similar procedure does not work as soon as we move to knowledge scenario 2. Once we know the semantic space, the user's undisclosed interpretation of the images is open to question. A second problem concerns the unknown query. The subset of images selected, i, will be a subset of $h^*(q)$ for a great many queries q and arrows h. How does the retrieval system know which of these is intended by the user? Clearly it cannot. What then do the images it retrieves represent? A third difficulty concerns the existence of the arrow σ. In most cases, it is obvious that such an arrow does *not* exist, in which case equation 1 will be incorrect: the method cannot produce the correct results. What then are we to make of the claims of successful retrieval reported in the literature? We leave further consideration of these issues until we discuss evaluation in section 2.3.5.

2.2 Evaluation Contexts

We move on from the structure of the retrieval system and the difficulties involved in its construction, to consider the evaluation of such systems. The evaluation of any system takes the form of a comparison between two arrows with common domain D and co-domain C: a 'reference arrow', which describes the ideal behaviour, and a 'test arrow'. We will call the co-domain C the 'evaluation context'. We introduce a probability measure μ on D and a 'loss function' $\rho : X \times Y \to \mathbb{R}$, where $X, Y \in \left\{C, 2^C\right\}$. Both these quantities are to be decided by the evaluator. In our case, the measure on D might correspond to the frequency with which certain queries are put to the system. The comparison between two arrows a and b then takes the form

$$\Upsilon(a, b) = \int_D d\mu(d)\, \rho(a(d), b(d)) \tag{2}$$

where a is the reference arrow, which in the case of retrieval will always involve h, and b is the test arrow, which should not involve h. (The integral includes the case in which D is discrete.) The value of Υ is thus a measure of how well/badly we are doing by using b instead of a. A score of 0 would indicate perfection: the arrows a and b are the same for the purposes in which we are interested. (Note that this does not necessarily mean that $a = b$.) This generic situation is illustrated in figure 2.2.

To specify an evaluation method it is necessary to specify C, D, a and b, but in the case of retrieval systems, the structure of figure 2 means that a and b are specified once C and D are given, and in fact there is always an obvious choice of D also. We can therefore concentrate on the evaluation contexts. There are three of them: the 'image context' (which we will also call the 'retrieval context'), the 'indexing context' and the 'semantic context'. Whether a given evaluation context can be used depends on which of the arrows in figure 2 are available, since computation of $\Upsilon(a, b)$ is clearly impossible if we cannot compute a and b. We will thus see that different evaluation contexts are appropriate in different knowledge scenarios. We now define the three evaluation contexts in more detail.

2.2.1 Semantic Context: $C = S$

The natural choice for D is I, so that the arrows being compared are h and $\sigma\pi$. The nature of ρ is likely to be highly application-dependent, but at least in this scenario we know what the points of C mean, and can define ρ accordingly. The measure μ could be based on the frequency with which particular images have been retrieved in the past or, more simply, could just be uniform.

2.2.2 Retrieval Context: $C = I$

This is the most familiar evaluation context, and the one most often evoked in the literature, presumably because it seems to offer the most direct connection to the 'average relevant user' and to retrieval performance. It is rarely used in the complete form presented here, using the semantic space as domain. More often, a version appropriate to the assumptions of query by example is used. We will discuss this further in section 2.3.5.

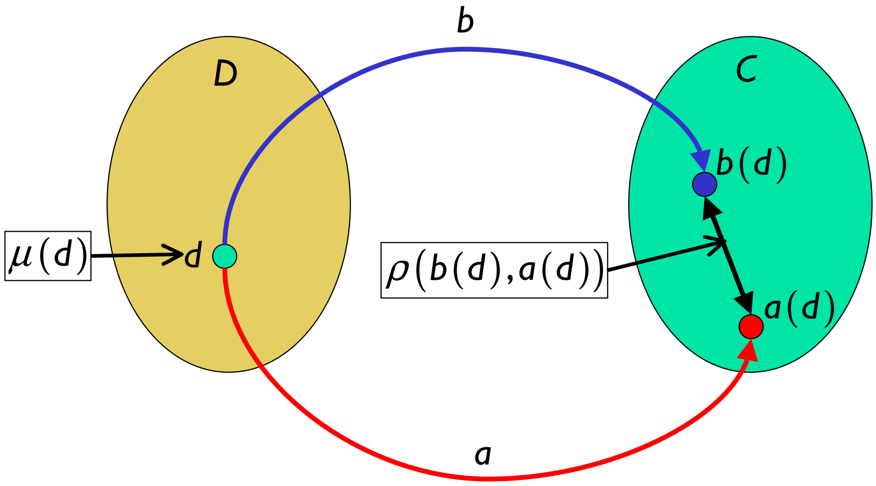


Figure 3: The structure of a generic evaluation. The arrow a is the reference arrow, giving the 'correct' value of the map between the domain D and the co-domain, or evaluation context, C. The arrow b is the test arrow, the arrow being evaluated. In the case shown, the arrows do not agree when evaluated on the element d of D. The difference between the two values is given by the loss function ρ on $C \times C$. The elements of D may be of differing importance. This information is contained in the measure μ on D. The score Υ is the average over D, in the measure μ, of the value of ρ.

The natural choice of domain is S, and the arrows that are compared are h^* and $\pi^*\sigma^*$. These map points of S to subsets of I, so that we must define a loss function on $2^I \times 2^I$. A typical choice of loss function is the "recall":

$$\rho(A, B) = \frac{|A \cap B|}{|A|} \tag{3}$$

It is the fraction of the images that the 'average relevant user' would have retrieved, that are retrieved by the system. Based on another obvious normalisation, one can also define the "precision", given by

$$\rho(A, B) = \frac{|A \cap B|}{|B|} \tag{4}$$

It is the fraction of the images retrieved by the system that the'average relevant user' would have retrieved also.

A natural choice for μ in this evaluation context, is the frequency with which queries are put to the system.

2.2.3 Indexing Context: $C = J$

The natural domain is I, and the arrows compared are π and $\sigma^* h$. The domain of ρ here is $2^J \times J$, but the choice of a loss function is far from obvious. The measure μ can take the same form as in the semantic context.

The main advantage of this evaluation context is that we do not need separate and explicit characterizations of h and σ, since only the combination $\sigma^* h$ appears in equation 2. We must however make an assumption. In order to compare the performance of different methods, we need a common J. This limits the range of applicability of the indexing context to the comparison of systems that share, at least to some degree, the same J. Often we can generate a common J by removing structure. Suppose for example that the index spaces of various segmentation algorithms consist of partitions of the image domain labelled by the values of different features. A direct comparison is impossible, but by keeping only the common structure (the image domain partitions), a comparison is enabled. The effect of this is to 'coarsen' the semantics we can hope to capture, since it will group each index space into equivalence classes.

2.3 Evaluation Contexts and Knowledge Scenarios

Having defined the evaluation contexts, we can now look more closely at when they can be applied, and in particular, in which knowledge scenarios they are relevant. It will turn out that lack of explicit knowledge of certain maps need not hinder us if we can define them implicitly. We look at the knowledge scenarios one by one, treating query by example separately, since it has properties peculiar to the assumptions used in its definition.

2.3.1 Knowledge Scenario 4

In this case, the natural evaluation context to use is the semantic context. All the arrows are defined, and we need only define a loss function on S. Since by assumption we can construct arrows $\sigma\pi$ that come close to h, we can expect reasonably high scores in our evaluations. In addition, the other contexts are guaranteed to give good results if we have an image processing method $\sigma\pi$ that duplicates h. Thus the retrieval context, although apparently well-adapted to this situation, is not really needed, and indeed is harder to use: it is more difficult to define meaningful loss functions on I than on S.

2.3.2 Knowledge Scenario 3

In this case, we cannot use the semantic context, or rather use of it is meaningless. The arrows $\sigma\pi$ that we know how to build come nowhere near duplicating h, so that claiming victory for one method over another is really beside the point. The same consideration applies to the retrieval context. In the case of the indexing context, one might hope that we could define an arrow $\sigma^* h$ without explicitly defining σ, but the fact that we already know h does not allow us the freedom to do so.

Thus knowledge scenario 3 turns out to be impossible to evaluate. This highlights the somewhat dubious nature of evaluation in the scenarios to come, in which we have even less knowledge. The lack of constraint allows us to make progress by making simplifying assumptions, but knowledge scenario 3 makes it clear that we should be wary of drawing hasty conclusions from any apparent success such evaluations may produce.

2.3.3 Knowledge Scenario 2

We cannot use the semantic and retrieval contexts since these rely on knowledge of h, which we lack. Similarly to knowledge scenario 3, we cannot make assumptions about $\sigma^* h$, since although we do not know h, we do know S. Again evaluation is not possible.

2.3.4 Knowledge Scenario 1

As in the previous scenario, the semantic and retrieval contexts are eliminated. The indexing context however is not eliminated so easily. As we have stated, in the indexing context we do not need the semantic space explicitly, nor do we need explicit characterizations of the individual arrows h and σ. We need only the combination $\sigma^* h$. Since there are now no constraints on the individual arrows making up the combination, we are freer to try to characterize this arrow in another way. The natural way to do this is through the use of human subjects. This is a subtle point: we are saying that although it is not possible to characterize the semantics of images directly, nevertheless we can gain access to some knowledge about those semantics by looking instead at the results in J that might generate those semantics correctly.

How should we set about using human subjects to characterize this arrow? Clearly human image understanding does not generate points in or subsets of J. We cannot therefore ask human subjects to tell us their interpretations and use these as the explicit characterization of the arrow h, as we might do if we had a well-defined semantic space. Indeed the semantic space is left unspecified. We can however ask human subjects to evaluate directly the points in J that are generated by the arrow π, thus characterizing h implicitly. Absolute scoring of individual arrows will not do, since the meaning of the absolute scores will be very unclear, but a comparison of the outputs of the different π arrows stemming from different methods is nevertheless possible. The subject can be asked which of the two representations generated by two different methods is more meaningful. The most interesting result of such an evaluation is the existence or not of consensus among subjects. Its existence indicates that there may be an underlying 'fundamental' image semantics to which we can gain access via such experiments. We discuss this methodology further when we discuss the BAL dataset in section 2.4.2.

2.3.5 Query By Example

We note first that there is only one arrow so far in query by example, $\pi^* \pi_*$, and that consequently there is nothing to evaluate. In order to proceed further, a second, reference arrow is needed. To this end, an arrow $\hat{h}$ is introduced, that notionally maps I to some semantics $\hat{S}$. (Neither $\hat{h}$ nor $\hat{S}$ are *a priori* the same as the quantities h and σ that are supposed to generate the 'example' images in the retrieval process itself.) The arrow $\hat{h}$ is not described directly, since to do so would require a definition of $\hat{S}$. Only the combination $\hat{h}^* \hat{h}_*$ is described, by giving the partition of I into equivalence classes sharing equal values of $\hat{h}$. The equivalence class of images to which a given image belongs is known as the set of images "relevant" to the given image.

The introduction of $\hat{h}$ enables the comparison of the arrows $\hat{h}^*\hat{h}_*$ and $\pi^*\pi_*$, the arrow defining retrieval in query by example, since both map 2^I to itself. One can use appropriately adapted versions of equations 3 and 4 to compare the number of "relevant" images retrieved to the number of "relevant" images in the database or to the number of retrieved images.

The way this is done in practice is the following. The database of images is divided into groups, supposed to represent the arrow $\hat{h}^*\hat{h}_*$. These groups are typically based on the 'generic name' of the 'most prominent object' in the image. To test the retrieval abilities of the system, an image or set of images i is pulled back by the arrow $\pi^*\pi_*$, giving a set of retrieved images, which are then compared to the set of images "relevant" to i. The results of these tests are sometimes quite remarkable. Recall and precision values above 90% are not at all unusual. Are we really this good at content-based image retrieval?

What does it mean that the images retrieved and the relevance classes into which I is divided agree so closely? In its raw form, it means that the arrow π has managed to duplicate the grouping of I into equivalence classes. In itself, this is not that impressive of course. Given enough parameters, any classification can be duplicated. The inference from the results, given the grouping of I, is however closer to the following: "An image of a horse was used as a query, and the retrieved images consisted of almost all the horses in the database and very little else. Thus the method is capturing the image semantics.". Let us analyse this statement.

The first point to note is that h and $\hat{h}$ are not necessarily the same. If they are not the same, then we would not expect retrieval based on $\pi^*\pi_*$ to agree with retrieval based on $\hat{h}^*\hat{h}_*$. Thus, while we may choose to imagine that the user was actually looking for horses, he may have been looking for images with any of a broad range of other interpretations. The user can change his mind about his interpretation at will, while still using the same set of 'example' images. Thus the fact that the retrieved set of images consists of horses may or may not be a success, depending on whether h and $\hat{h}$ are equal. Calling "obvious" the interpretation that renders the retrieval a success, does not solve this problem.

The second point concerns the existence or not of an arrow σ such that $h = \sigma\pi$. (We now assume that $\hat{h} = h$.) In the case that such an arrow exists, then, as shown in section 2.1.5, we will have that $h^* = \pi^*\pi_*h^*$. This means that precision will be 100%. If further, we have that σ is a bijection, then recall will be 100% also. The values of recall and precision reported in the literature suggest that this situation is close to being reached. This means that π divides the space of images into the same equivalence classes as h. Since h is never specified, it is of course unclear what this actually means. The clear implication however is that π is actually classifying images into the 'generic classes' of the 'most prominent object' in the image, that is, horses, cars,.... This is remarkable in methods that contain no models of these objects, and which sometimes use the crudest of global features. In fact, it is apparent that π is achieving no such thing. What then to make of the success of the retrieval experiments? Clearly, I must possess a remarkable structure. Firstly, the 'generic name' of the 'most prominent object' is closely correlated with the low-level features typically involved in π, and secondly, the images are well-clustered in J. Since neither is true in general, we are forced to conclude that the I being used to test the methods is very special, and that little can be made of the results reported.

2.4 Specific Application Domains

We turn now to a consideration of the application domains that we described in section 1.1: the IGN and BAL datasets, and how they fit into the above analysis. In section 3, we describe the implementation of this analysis on the BAL dataset. Evaluation using the IGN dataset is in progress: the situation is described in section 2.4.1.

2.4.1 IGN Dataset

For the IGN dataset, the semantic space S is given by the conjunction of statements such as those mentioned in section 2.3.4. One could develop a more complex semantics for the IGN images, involving recognizing more subtle variations in land use, or even individual objects such as people and cars were the resolution higher, but for many situations the simpler semantics is adequate. In addition, the actual land use is known, having been compiled from existing maps and field studies. Thus h is characterisable in the form of a ground truth land-use map for each aerial image in the database. In addition, the semantic space is such that we have a reasonable chance of being able to construct the semantic map σ, or equivalently of extending π all the way to S. The nature of the images in the IGN dataset creates this possibility: at the resolution of the images, the scene itself is more or less a flat two-dimensional surface, with texture 'painted' onto it, and in addition the different types of land use seem characterisable in terms of texture descriptors and other low-level image features. We are thus in knowledge scenario 4.

The way forward now depends on exactly what we want to test. If we want to test the performance of the image processing methods with respect to retrieval, then, according to the above analysis, the ideal context is the semantic context. Each method will produce a partition of the image domain labelled with various land uses, and we can then compare, based on a metric such as the percentage of land area misclassified, the performance of the methods.

Another possibility is that we wish to test segmentation methods in a way that is independent of classification. In this case, we are interested only in the partitions and not in the labels attached to them, so that we can simplify S by 'forgetting' the labels, and compare partitions directly.

2.4.2 BAL Dataset

For the BAL dataset, the semantic space seems unbounded. The Bridgeman Art Library has to deal with queries of a very abstract nature, whose relation to image properties is extremely complicated, involving a great deal of cultural knowledge. In addition, individuals may not be clear about their own interpretation, and it is almost certain that there will not be consensus over some of the statements one might like to make about the images.

We can however simplify this situation somewhat. Whatever the nature of the statements we wish to make about the images, it seems clear that they will require as a necessary, although by no means sufficient input, the identification of the 'principal objects' in the image.[6] We can therefore reduce our semantics somewhat by restricting ourselves to disjunctions and conjunctions of statements of a form rather similar to those used for the IGN images: "Such-and-such region of the image contains such-and-such (named) object". Unfortunately, this is still too broad. Statements about the BAL images contain a far larger set of objects than the IGN images, so many in fact that it is not feasible to list them all. We could give a fixed list of objects and define a semantics in those terms, but this is far too restrictive in practice. The absence of any well-defined semantic space puts us in knowledge scenario 1. The semantic and retrieval contexts are thus immediately ruled out.

At this point, we could try to invoke the assumptions of query by example, and at the same time classify the images in the BAL dataset into "relevance" classes as described above. The drastic nature of these assumptions is exposed once we start thinking about applying them to image and semantic spaces as complex as those of the Bridgeman Art Library. Most images from the BAL dataset do not clearly specify any query, and any attempt to categorize the images into "relevance" classes for the purpose of evaluation will fail also.

Instead, we turn to the indexing context, which, free as it is from the need explicitly to define S and h, has not yet been ruled out. We assume a working hypothesis: that there does exist human consensus about what might be called 'fundamental' image segmentations. (As recent work has shown [12], consensus may well exist at least for limited classes of images.) Under this assumption, it is reasonable to model h as a map. We then proceed as follows. In order to compare a number of segmentation procedures, they must share a common J. This we ensure by defining J as a space of (unlabelled) partitions of the image domain. The above assumption amounts to assuming that for each image, there is human consensus about a semantic interpretation that includes as part of its definition an image domain partition. We can thus ask individuals to 'score' the output of various segmentation algorithms by comparison with the original image and, in practice, to avoid the arbitrariness involved in an absolute scoring system, by comparison with each other.

Note that what is being tested is not simply the performance of different methods, but the very existence of a consensus about the interpretation of the images involved. In the absence of an explicit h, this is far from obvious. Indeed the existence or non-existence of a consensus is a question arguably more interesting than the results of the evaluations themselves.

2.5 Conclusions of Methodological Analysis

The ease of application of query by example, coupled with the notion that there is something 'special' about image data, seems to have created the impression that for image retrieval it may in principle be preferable to query by text. It is undeniable that there is something special about image data, at least when compared to text or speech retrieval. The segmentation of sound into recognized words leaves one with text, which is composed of atoms that exist already at the semantic level: one does not need to 'name' words. The number of atoms is limited and they are known *a priori*. In image

[6]The phrase 'principal objects' is not well-defined. However, for evaluation using the indexing context this is not important, since we are never required to identify such objects explicitly (unlike in the typical evaluation of query by example discussed above). This is the force of the argument in the text that we do not need explicit and separate characterizations of the arrows h and σ^* in the indexing context, but only a characterization of the combination $\sigma^* h$. Effectively, the phrase is used merely as a justification for talking about segmentation algorithms as opposed to, say, global colour histograms.

understanding, the number of semantic atoms is vastly greater if they exist at all, and the direct correspondence of segmented regions (for example) with semantics is less clear. Coupled with higher dimensionality, which allows geometry to intrude, and the projective nature of image formation, we see that image understanding is vastly harder than speech processing. We reject the notion however that these differences require a qualitatively different approach to image database retrieval. This is for the simple reason that semantics seems to us inherently linguistic. This is supported for example by the psychophysical experiments performed using the *PicHunter* system [5]. Query by example in which the meaning of the example is not clarified both to the user and to the system by linguistic cues that specify a query in a well-defined semantic space, simply results in confusion, as the above analysis shows. We believe that it cannot be a substitute for the characterization of a meaningful semantic space, an interpretation arrow h, and a genuine image processing arrow $\sigma\pi$ from I to S. Our current inability to construct such arrows in many of the cases of interest should not be disguised by lack of methodological clarity.

The evaluation method that combines "relevance" classes with QBE suffers from a number of serious drawbacks, not the least of which is the appearance of objective evaluation without the substance. 'Success' in such evaluations is less an expression of the ability of the image processing system involved, and more a statement about the distribution of the images in the database. Image retrieval systems *can* be evaluated within the semantic and retrieval contexts, but only if we have a characterization of S and h. In the absence of such characterizations, we are forced to move to the indexing context, and to perform psychological experiments with human subjects in order to evaluate systems.

Many of the above considerations apply directly to the evaluation of image processing methods in general. The reason for this is clear: image retrieval is in essence pullback by the image interpretation arrow $\sigma\pi$. The accuracy of the retrieval is entirely dependent on the accuracy of this interpretation. For cases in which no well-defined semantics is available, the only available evaluation method for image processing systems is the psychovisual one proposed above for the BAL dataset. The process involved in such experiments is similar to that used in the 'eye of the beholder' method of evaluation that is all too common in image processing. The difference is that properly designed experiments take into account a large number of different images and a range of different users in order to test the idea of a consensus and produce an evaluation if such a consensus exists.

3 Knowledge Scenario 1: the BAL Dataset

In this section, we describe in detail the implementation of the ideas discussed in section 2.4.2. There we applied the methodological analysis to the BAL dataset and its associated application, and noted that we landed squarely in knowledge scenario 1. Our analysis then leads us to use the indexing context as the appropriate evaluation context, and psychovisual tests to characterize the arrow $\sigma^* h$. We also concluded that even so, developing an evaluation method is difficult. We decided that a necessary (but not sufficient) condition for capturing the semantics of the BAL dataset would be to identify the 'principal objects' in the image, and we therefore turned to segmentation algorithms as a natural way to proceed. Certainly segmentation algorithms are not the only way to attempt to identify objects, so why focus on them? The reason is that the evaluation of segmentation algorithms is often as methodologically difficult as that of retrieval systems, and for the same reason: the absence of a well-defined S. As argued in the previous section, a well-defined evaluation of a retrieval system is essentially equivalent to a well-defined evaluation of the image processing arrow it contains, so that evaluation of retrieval systems that use segmentation algorithms can equally be considered as an evaluation of the segmentation algorithms themselves. We thus focus on segmentation algorithms due to their intrinsic interest.

The rest of this section is structured as follows.

Section 3.1 We describe the evaluation method in outline, and discuss possible alternatives.

Section 3.2 We give a short description of each of the segmentation schemes tested.

Section 3.3 We describe the experimental procedure in detail.

Section 3.4 We discuss and analyse the results obtained from the evaluation.

3.1 Summary of Method and Alternatives

In section 2.4.2, we gave an argument for conducting psychovisual tests that compared the results of different segmentation schemes[7] to one another and to the original image. Here we expand this argument and consider the alternatives.

In the psychovisual tests used here, human subjects are asked to choose between the segmentations resulting from different schemes. They are presented with the original image and two segmentations, one from each of two schemes applied to the original image. Each segmentation is presented in two different ways to aid the subject. In addition to the binary variable indicating the subject's choice between the two segmentation schemes, the time taken for the subject to reach a decision is recorded. This process is repeated on different images and with different pairs of schemes. Analysis of the results, described in section 3.4, then leads to a ranking of the schemes.

As we stressed in the conclusion to section 2, the emergence or not of a consensus is perhaps the most interesting result of an investigation such as that carried out here. If no consensus emerges, then it is hard to claim that the notion of a segmentation in the absence of an explicit semantic space is amenable to evaluation at all. It is a first check on the existence of a consensus that the pairwise

[7]Hereafter we use the word 'scheme' to refer to a segmentation algorithm or method.

ranking generated by the tests be consistent with an overall total order on the schemes. This is not necessary: cycles may exist in the pairwise ranking rendering this impossible. We discuss this point in more detail in section 3.4.

What about alternative testing methods? One possibility is to show subjects the result of a segmentation scheme and the original image and ask them to assign an absolute score to the segmentation. We dismissed this approach because of the difficulty of ensuring consistency in the meaning of the absolute scores across different subjects and across different images.

Another possibility is put into practice in [12], who consider the evaluation of segmentation schemes *per se*. They too treat a situation in which the semantic space is ill-defined by turning to human subjects but, in contrast to the approach advocated here, they ask subjects to segment images by hand. If a reasonable consensus emerges, the hand segmentations can be treated as ground truth, and compared directly to the outputs of segmentation schemes, as in knowledge scenario 4. We find a number of difficulties with this approach.

The first is that there may be consensus about several possible semantics for the images: the arrow h, viewed as a conditional probability, might be multi-modal. One is then faced with the difficulty of marginalising over the semantics. In contrast, using comparison of segmentations generated by the segmentation schemes implicitly marginalises over semantics by allowing the subject to use whatever semantics is appropriate.

The second problem is that the segmentation task may be too difficult if the images to be segmented are semantically sufficiently complex. Our trials with hand segmentations suggested that for the BAL images, subjects simply did not know what to do when faced with the segmentation task. Clearly this issue is closely related to the first difficulty, and indeed it seems intuitively likely that the existence of a multi-modal conditional probability for the semantics is the cause of the problems subjects experienced in manually segmenting the BAL images.

In [12], these problems did not seem to arise. The paper does find a degree of consensus across the image data set used, which is a subset of the Corel natural image dataset, although it is worth noting that consensus is judged by using a hierarchical measure that allows regions to be sub-divided. We suspect that the above problems are absent due to the nature of the images in the Corel dataset. Many of the images in the Corel dataset have a relatively simple structure: a small number of significant objects in the centre foreground, and a small number of easily separable background classes. The BAL images on the other hand are generally more complex.

The third difficulty concerns the loss function, ρ, on the space of segmentations. When we are in knowledge scenario 4, for example in the case of the IGN dataset, we are defining a metric on a given semantic space S. We thus know *a priori* the meaning of the segmentations, and the application itself may give us a loss function. For example, the identification of forest as agricultural field may be a relatively unimportant error, whereas the converse may be significant for the application at hand. Even in the absence of such a clear indication, the fact that the segmentations have meaning allows us to use our intuition in defining a loss function. In the case of knowledge scenario 1 and segmentations without explicit semantic significance, it is much harder to justify a choice of loss function. This does not mean that it is hard to define one; [12] define an ingenious metric that takes into account the assumed hierarchical nature of segmentations. It is less clear however that this choice corresponds to the metric one might develop if one had access to the unknown semantics of the users. Despite these reservations, the results in [12] are extremely interesting, and are complementary to the evaluation technique presented here.

3.2 Image Segmentation Schemes and Dataset

Six schemes were made available for evaluation, of which five came from the members of the MOUMIR project and one from outside (Blobworld). Details of the features used, the models and the algorithms can be found in the cited papers.

- *Multiscale Image Segmentation (MIS)*: This scheme is outlined in [8]. It generates classes using a robust mean shift procedure, which operates on a 7 - dimensional joint spatio-feature space, containing 3 colour, 2 texture and 2 spatial feature components. The subsequent classification procedure consists of a Bayesian multiscale process which models the inherent uncertainty in the joint specification of class via a Markov Random Field model.

- *Blobworld*: The Blobworld scheme aims to transform images into a small set of regions which are coherent in colour and texture [3]. This is achieved by clustering pixels in a joint colour-texture-position feature space using the EM-algorithm.[8].

- *Iterated Conditional Modes* (ICM): The likelihood for this model uses an i.i.d. Gaussian model of pixel intensities, combined with a Potts-like prior. ICM is used to maximise the posterior probability, an initial configuration being created using k-means [17].

- *Learning Vector Quantization* (LVQ): The feature vectors used in the LVQ clustering algorithm consisted of the RGB colour values and the coordinates of each pixel. LVQ is a self-organizing neural network with a competitive learning law [11].

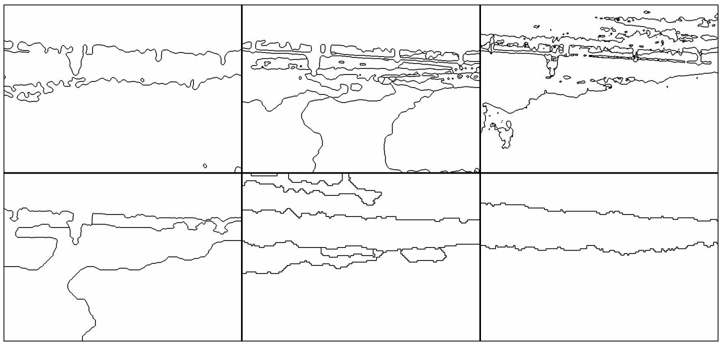
- *Double Markov Random Field* (DMRF): The double Markov random field assumes Gaussian MRF models for classes within the image, and that class labels follow a Potts model [14]. In this approach, the posterior distribution of class labels and all model parameters given observed intensities are simulated using a Markov chain Monte Carlo approach. The segmentation is taken to be the marginal posterior mode, where each pixel is classified to be that class that was sampled most often in the simulation.

- *Complex Wavelets and Hidden Markov Trees* (CHMT): Segmentation using Complex Wavelets and Markov Trees [20] is initialised using the mean shift procedure to generate classes. Then texture and colour models, based on hidden Markov trees of complex wavelet [10] and scaling function coefficients respectively, are trained. A segmentation is found by using maximum likelihood classification of the coefficients given the models [6].

An idea of the variation in the segmentations produced by the schemes that we evaluate can be obtained from figure 4(b), which shows the output of the six different segmentation schemes given the image in figure 4(a).

[8] It was impractical to alter any of the internal parameters of Blobworld .

(a)



(b)

Figure 4: 4(a) shows the variation in the segmentations of the image in 4(b) resulting from the six schemes evaluated (Image © Bridgeman Art Library).

3.3 Evaluation Method

The psychovisual test consisted of a series of 'trials' called an 'evaluation set', each of which asked a human user ('subject') to choose between the segmentations of a single image by two different schemes. Six schemes were compared to each other, thus giving 15 pairwise comparisons. The number of subjects was 14.

Previous work in the area of psychovisual testing [9] suggests that a 30-minute time limit should be placed on the length of the test. This is to guard against subject fatigue, which could influence the results in an unpredictable manner. During preliminary investigations of the testing process, we found that each trial took about 10 seconds. An evaluation set therefore consisted of 150 trials. The existence of 15 pairwise comparisons meant that each pair of schemes was compared over 10 trials.

Two subsets of 10 and 50 images respectively were selected at random from an initial set of 3000 images from the BAL dataset. Although the segmentation schemes are unsupervised, some minimal setting of parameters is required. This is to avoid drastically over- or under-segmenting the images, thus adding further difficulty to the evaluation problem. Thus, the first subset of 10 training images was used to fine-tune the parameters of the schemes. The fine-tuning was done by the authors of each scheme. Each image was accompanied by rough guidelines indicating the number of regions expected, and some idea as to the identity of the principal regions in each image. The number of regions ranged from 2–10. After the fine-tuning, the parameters were left untouched while the schemes segmented the remaining 50 test images. To perform ten trials with each pair of schemes required ten images per pair. These were found by sampling without replacement from the 50 test images, the sampling beginning again when all 50 images had been used. The resulting 150 trials were then randomly sequenced, and the two schemes in each trial assigned randomly to left or right for display purposes (see section 3.3.1). Once done, this same evaluation set was used for each subject.

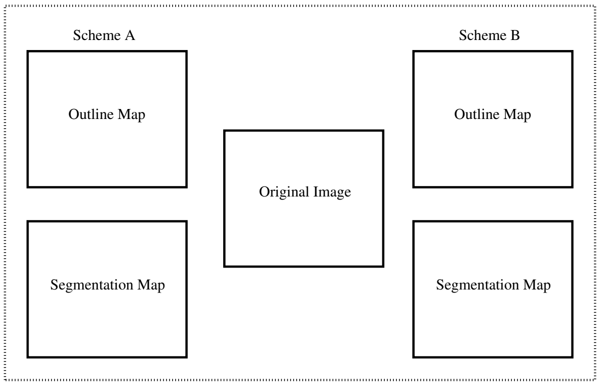
3.3.1 Trial

In each trial, five images were displayed. The original image was displayed centre screen, with, on either side of it, the segmentation results from the two schemes being compared in that trial. Each scheme's segmentation was presented to the subject using two different representations. We term these images the 'outline map' and the 'segmentation map'. The outline map shows the region boundaries superimposed on the original image. The segmentation map shows the regions themselves by colouring a region with the mean colour of the regions corresponding to its class. Figure 5 illustrates the layout of a trial.

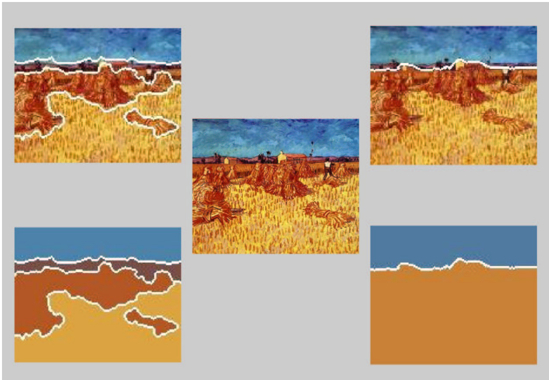
The reason for using two representations of a single segmentation result is the following. The outline map is designed to display the coincidence (or not) of region boundaries with features in the image, and the nature of the interior of each region. The segmentation map is designed to show the global layout of the regions and to show which regions are classified as belonging to the same class. It also presents a greatly simplified version of the original image.

The images in each trial were displayed in two stages. First, the original image was displayed on its own. The subject had an indefinite amount of time to look at this image and become familiar with it. When ready, the subject could click on the original image, at which point the other four, segmented images were displayed. When the subject had made a decision, the next original image was displayed, thus commencing the next trial. Figure 6 shows these two stages of a trial.

During each trial, two types of measurements were made of the subjects' response to the proposed segmentations: a 'soft score' and a 'hard score'. The soft score is based on the time taken for the subject to make their decision after the segmented images are displayed. The idea of displaying the original image first, and then of waiting until the subject is ready to proceed, was to reduce as far as possible the effects of image complexity on this soft score. In practice, for analysis, we use the reciprocal of the decision time, termed the 'speed'. The speed is signed, the sign of the speed indicating which of the two schemes the subject chose. If the subject was 'undecided', the speed was



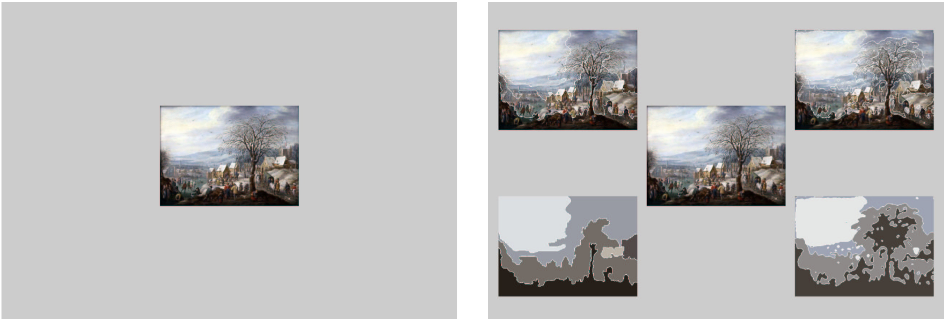
(a)



(b)

Figure 5: 5(a) is a schematic of the on-screen layout of a trial, while 5(b) contains a sample screen shot (Image © Bridgeman Art Library).

set to zero. The measurement of speed is intended to indicate how close in meaningfulness the segmentations from the two schemes are to each other (although not necessarily close geometrically), a slower decision indicating that the two segmentations were closer. This type of 'speed of response' measurement has proven to be an effective measure in other psychovisual tests related to detecting image compression artifacts [9]. The hard score consists simply of the sign of the speed, therefore indicating which scheme the subject chose, but not how quickly. These two measurements form the basis of the results analysis in section 3.4.



(a) (b)

Figure 6: This figure contains the two stages of a trial. Figure 6(a) shows the first stage of a trial, in which the original image is displayed to the subject until they are familiar with it. To proceed to the trial, the subject clicks on the original image. Figure 6(b) illustrates the second stage of a trial, in which the segmentations from the two schemes are represented by their outline and segmentation maps. This stage of the trial is timed, resulting in the soft score described in the text. (Image © Bridgeman Art Library)

3.3.2 Instructions

Prior to performing the psychovisual tests, all subjects were issued with a set of instructions. The instructions were designed to be minimal, in the sense of influencing the semantics the subject would use to understand the image as little as possible. The instructions read: *The pair of images to the left of the original image illustrates one way of splitting the original image into its most important pieces, while the pair of images to the right of the original image illustrates a second way. Decide which of the ways, left or right, of splitting the image into its most important pieces makes most sense to you*. Once the subject comes to a decision, using a mouse they click on either of the outline or segmentation maps of the chosen scheme. Subjects may also be *undecided* in their choice, in which case they click on the original image.

In order to allow subjects to familiarise themselves with the test, each subject underwent a short 'familiarisation session' prior to performing the test proper. The session consisted of 10 trials. The images used in the familiarisation session were not used again.

3.4 Results and Discussion

The goal in analysing the experimental measurements is to determine whether or not a clear and consistent ranking of the schemes exists, thus *evaluating* the schemes. To illustrate how we combine the measurements obtained from each subject over 150 trials we will focus initially on the 'soft' measurements.

	CHMT	ICM	DMRF	LVQ	Blob.	MIS	Overall Score
CHMT	–	-5.31	5.78	14.81	15.55	8.95	51.4
ICM	5.31	–	11.33	9.24	10.97	20.56	45.8
DMRF	-11.33	-5.78	–	6.92	31.73	0.46	22
LVQ	-9.24	-14.81	-6.92	–	6.89	35.12	11.03
Blob.	-10.97	-15.55	-31.73	-6.89	–	17.55	-47.59
MIS	-8.95	-20.56	-0.46	-35.12	-17.55	–	-82.64

Table 1: Each row of the table contains the soft pairwise scores for a scheme with corresponding the overall score in the rightmost column. For display purposes each score is multiplied by 1000. For example, the CHMT scheme scored 14.81 against LVQ. A positive score in a row is good for that scheme.

To achieve a ranking of the schemes, an overall 'score' is computed for each scheme. This score is determined by combining individual trial scores, v'. The trial scores are in turn computed from the speed of the subjects' responses. In order to compare the scores from different subjects, some model has to be given to account for the variation among subjects. In principle, this variation could be extremely complicated, in which case there is little hope of discovering a consensus. We make perhaps the simplest assumptions possible. The speed of response, $v(u,t)$, during a trial t with subject u is assumed to take the form:

$$v(u,t) = \alpha(u)d(t) + N \tag{5}$$

where: $d(t)$ is the 'true' measure of how much better one scheme was than the other at segmenting the image in trial t; $\alpha(u)$ is the subject's 'speed coefficient'; and N is Gaussian noise of zero mean and unknown but constant variance. The dependence on the subject is removed by normalising the scores:

$$v'(u,t) = \frac{v(u,t)}{\|v\|(u)} \tag{6}$$

where $\|\cdot\|$ is the Euclidean norm over the trials. To compute a score for each scheme we examine each trial in turn and add a score of $+|v'|$ to the 'winning' scheme, $-|v'|$ to the 'losing' scheme. Note that v' is 0 if the subject was undecided. These scores are subsequently averaged over all trials and subjects, thus giving an overall score, and hence a ranking of the schemes. A second possible ranking is obtained by examining the pairwise scores. The pairwise scores are computed by averaging over the relevant v' values obtained when comparing two particular schemes. The pairwise and overall scores due to the soft measurements, in overall rank order, are shown in table 1 and graphically in figure 8(a). A similar approach is taken to the hard measurements, however, in this case we assign $v' = 1$ and proceed as previously discussed. The hard scores, in rank order, are shown in table 2 and graphically in figure 8(b).

Given the results, the following natural questions arise: first, what are the rankings associated with the pairwise and overall scores and do they agree with each other? Second, do the rankings obtained from hard and soft scores agree?

	CHMT	ICM	DMRF	LVQ	Blob.	MIS	Overall Score
CHMT	–	-0.04	0.1	0.14	0.19	0.07	0.5
ICM	0.04	–	0.13	0.09	0.09	0.11	0.42
DMRF	-0.13	-0.1	–	0.14	0.27	0.04	0.22
LVQ	-0.09	-0.14	-0.14	–	0.11	0.34	0.08
Blob.	-0.09	-0.19	-0.27	-0.11	–	0.11	-0.55
MIS	-0.07	-0.11	-0.04	-0.34	-0.11	–	-0.67

Table 2: The table shows the hard pairwise and ranking scores. Each row of the table contains the scores of that scheme against the others.

The first part of the first question conceals a hidden complexity. Given a pairwise ranking of the schemes, it is not necessarily the case that there exists a total order consistent with it: there may be cycles in the pairwise ordering. The existence of a total order consistent with the pairwise ordering gives a first indication that a consensus exists. Figure 7 shows a graph in which a thin arrow from vertex A to vertex B indicates that the scheme associated with vertex A performed better than that associated with vertex B in the pairwise scores. The thick arrows show the total order that results from the pairwise orderings. (Note that this is not necessarily the same as the total order resulting from the overall scores.) The diagram illustrates how unlikely it is *a priori* that such a pairwise assignment would lead to a total order. For six schemes, there are approximately 45 possible pairwise orderings for each possible total ordering.

The second part of the first question is answered as follows. The ranking due to the overall scores is CHMT $>$ ICM $>$ DMRF $>$ LVQ $>$ Blobworld $>$ MIS. The ranking due to the pairwise scores is the same except that the order of ICM and CHMT is reversed at the top of the list. This difference is probably due to the similarity in performance of the two schemes. In answer to the second question, the pairwise rankings produced by the hard and soft scores agree, as do the overall rankings produced by the hard and soft scores.

Two other questions then arise: how close are the results to an 'ideal' set and, conversely, how close are the results to those that would be obtained from subjects who randomly chose 'winning' schemes? Let us first describe an ideal set of results and illustrate how we propose to measure the consistency of the results and compare the results obtained. We propose that an ideal system should produce a set of consistent and coherent pairwise and overall rankings. In addition, the consistency should not only apply to the rankings but also to the scores obtained; for example, the 'best' scheme should have its largest pairwise score when it is compared to the 'worst' scheme. Were such an ideal system to exist, then a visual representation, for either hard or soft scores, would be somewhat like figure 8. To determine how close the results obtained in practice are to this ideal set, we propose to use "Leave-One-Out" analysis using correlation. Given a set of scores, S, (such as in tables 1 or 2), we proceed as follows.

1. For each of the M schemes in turn, we remove from S the row vector x_m corresponding to scheme $m \in M$. This gives a reduced set of scores, S_m. The row vector has $(M-1)$ entries; the diagonal score is omitted.

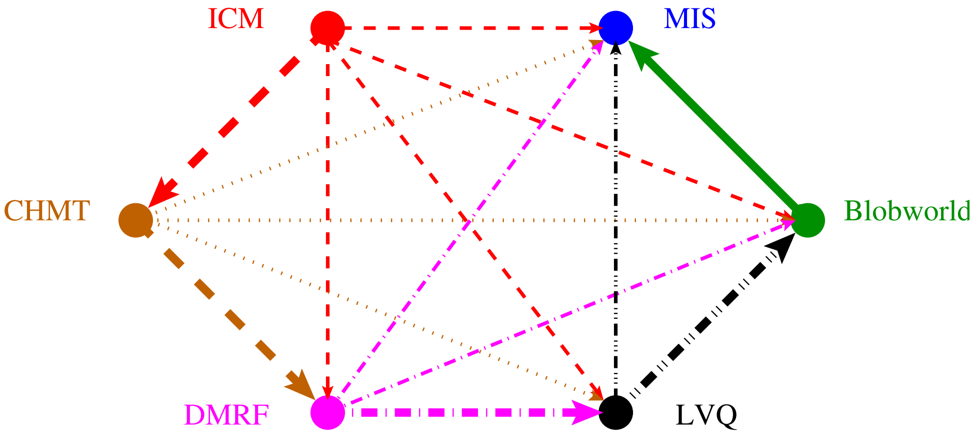


Figure 7: A graph illustrating the pairwise ordering arising from both the hard and soft scores, and the resulting total ordering of the schemes. (Note that this ordering is not necessarily the same as the total ordering derived from the overall scores.) There is a thin arrow from vertex A to vertex B if the scheme associated with vertex A performed better than that associated with vertex B in the pairwise ordering. The resulting total order is shown using thick arrows.

2. We then perform a summation across the rows of S_m, giving a column vector, denoted y_m, with $(M-1)$ elements. Both x_m and y_m are shifted so as to have zero mean.

3. The correlation between x_m and y_m is used to describe the similarity and consistency in scoring. To remove scale dependence the result is divided by the product of the Euclidean norms of the vectors. This gives us a score:

$$\psi_m = \frac{x_m \cdot y_m}{\|x_m\|\|y_m\|} \tag{7}$$

   where $\|\cdot\|$ is the Euclidean norm of the vector, and $\cdot$ is the Euclidean inner product. It can be shown empirically that in the ideal case $\psi_m = 1$.

4. A measure of how consistent the left-out pairwise results are with the overall score is then determined by averaging the ψ_m's.

This analysis is repeated for both sets of scores. To determine how likely or unlikely our measured results were, 10,000 randomly generated antisymmetric matrices were analysed using the same analysis. The results of this experiment are displayed in figure 8. We see from the data that the soft and hard scores obtained are estimated to have a 19% and 22% likelihood, respectively, of being produced through random selections.

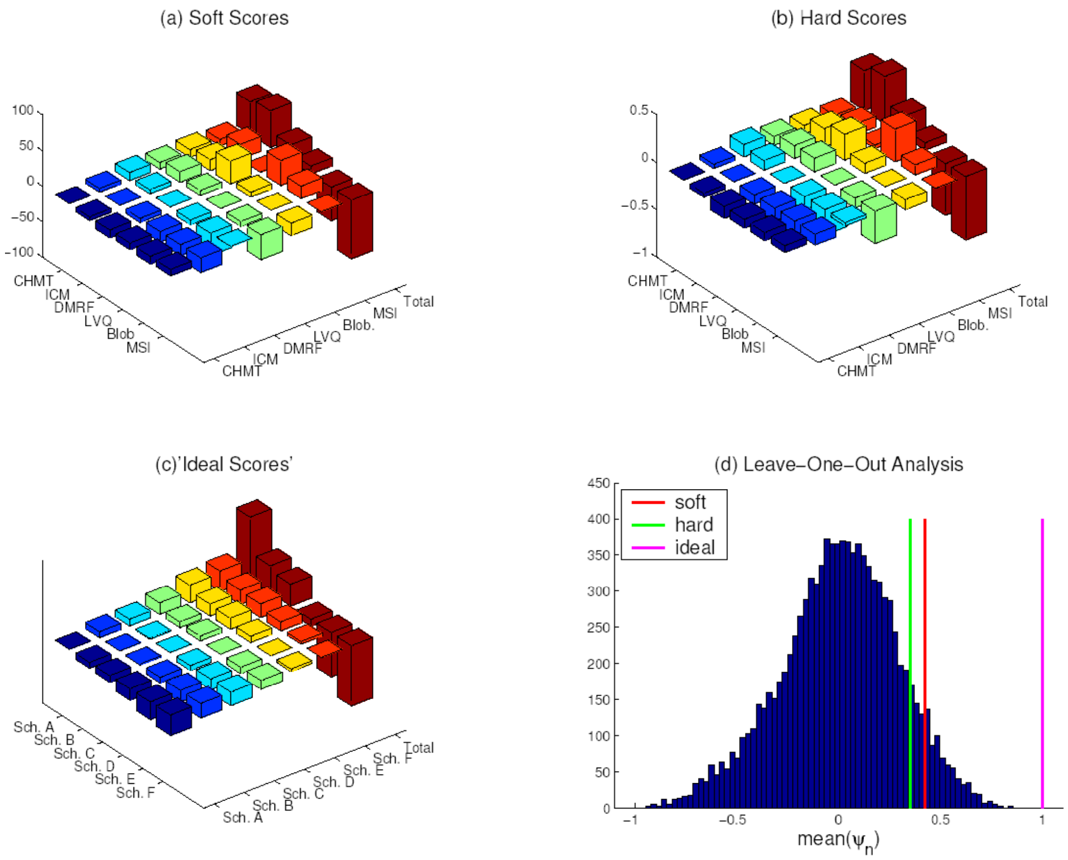


Figure 8: Graphical representation of the soft, hard and ideal scores in (a), (b) and (c) respectively. (d) Contains a histogram (in blue) of the mean ψ values obtained through random realisations of an antisymmetric scores matrix. An ideal system produces a value of 1, while the soft and hard scores produce, 0.422 and 0.351 respectively.

The question of what influences the final ranking is intriguing. ICM has the simplest feature set of any of the algorithms tested, eschewing even the use of colour, while CHMT has one of the most complex, training hidden Markov tree models of wavelets for texture, and using scaling coefficients for colour. Yet these schemes perform very similarly. It seems likely that the sophistication of the features and model in the CHMT scheme, which takes into account interscale dependencies between wavelet coefficients, explains its success. In the case of ICM, it might be that the use of greyscale intensity differences as the sole feature means that region boundaries coincide more precisely with significant boundaries in the image, thus helping to compensate for the lack of sophistication in its features.

Perhaps the closest direct comparison is between ICM and LVQ. Both use simple features: ICM uses pixel intensity while LVQ uses pixel colour. This may suggest that intensity gradient is a better indication than colour gradient of semantic boundaries in the image. These two models also differ in the way they take into account region geometry. LVQ clusters pixels using their coordinates as another feature, while ICM applies a Potts prior favouring smaller total boundary length.

Connected to this difference is the possibility that the number of regions (as opposed to classes) plays an important rôle. At least as limiting cases (one region, or a very large number), it is clear that this is a significant factor. LVQ tended to produce a large number of regions, whereas ICM produced a reasonable number (say ~ 5). This is not a consistent interpretation of the results however, as Blobworld also produces a reasonable number of regions. Perhaps Blobworld's use of an *a priori* structure for these regions hampers its performance on the images in the BAL database. Blobworld was originally applied to natural images often possessing a single dominant object in the centre foreground.

There are clearly many other possibilities that might help explain the total order that was found in our experiments. However, to confirm any particular proposition would require a great deal of further experimentation, with many more users and images. While seeking such an explanation of the data, it is worth bearing in mind the following. We are seeking a function of segmentations s and images i, $E(s,i)$ that explains the data in the sense that it increases as the rank of the segmentation scheme in the total order increases. If we possessed such a function, then it would in itself constitute a segmentation scheme: $s^*(i) = \arg\max_s E(s,i)$. Thus the correct explanation of the data would be the E that duplicates the hypothesized 'fundamental' human segmentation. Finding the correct E is therefore not easy. It is possible that for a limited dataset and number of users, one could find an E that duplicated the ranking of the data, but the status of such an E would always be somewhat suspect because of the above argument. In our case, it seems that the natural choice of E would be the optimization criterion of the ICM scheme, or in other words a Potts prior energy added to a Gaussian noise term in the intensity with mean and variance that depend on the class. We are in the process of investigating other measures to see if a definitive picture emerges.

In summary, the results of our experiments are consistent with a total ordering on the six schemes tested, and this ordering is essentially unaltered by several means of analysing the data. The results are far from chance levels. This consistency suggests that there is a degree of consensus among human subjects: they do perceive images as broken up into regions in a consistent way. The results of our study are inconclusive about the reasons for the ordering we obtain. The two most successful schemes use very different features and models. Further testing and analysis is necessary to determine if there are commonalities linking the successful segmentations. As argued in section 2, an image segmentation scheme that closely mimics the human interpretation of semantic content is in a better position to attempt retrieval of that content within a CBIR setting than one that does not. Work integrating the CHMT scheme into a CBIR framework is currently under way. The results will be published at a later date.

Acknowledgements The authors thank the following people and institutions: the Bridgeman Art Library, the Institut Géographique National, and the Institut d'Aménagement et d'Urbanisme de la Région d'Ile-de-France for their permission to use the images in the report; Dr. Erinija Pranckeviciene from the Artificial Intelligence and Information Analysis Laboratory of the Aristotle University of Thessaloniki, Greece, and Dr. Simon Wilson from the Statistics Department of Trinity College, Ireland, who together provided three of the segmentation schemes used in the evaluation process; the members of the Signal Processing Laboratory of the Cambridge University Engineering Department and Project Ariana (CNRS/INRIA/UNSA) at INRIA Sophia Antipolis for agreeing to act as subjects. This work was supported by European project MOUMIR, HP-99-108 (`http://www.moumir.org`).

References

[1] S. Ardizzoni, I. Bartolini, and M. Patella, *Windsurf: Region-based image retrieval using wavelets*, Proc. DEXA Workshop, 1999, pp. 167–173.

[2] R. Brunelli and O. Mich, *Image retrieval by examples*, IEEE Trans. Multimedia **2** (2000), no. 3, 164–171.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik, *Blobworld: Color- and texture-based image segmentation using EM and its application to image querying and classification*, IEEE Trans. Patt. Anal. Mach. Intell. **24** (2002), no. 8, 1026–1038.

[4] V. Chalana and Y. Kim, *A methodology for evaluation of boundary dectection algorithms on medical images*, IEEE Trans. Med. Imag. **16** (1997), 642–652.

[5] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, *The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments*, IEEE Trans. Im. Proc. **9** (2000), no. 1, 20–37.

[6] M. Crouse, R. Nowak, and R. Baraniuk, *Wavelet-based statistical signal processing using hidden markov models.*, IEEE Trans.Sig. Proc. **46** (1998), 886–902.

[7] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher, *An experimental comparison of range image segmentation algorithms*, IEEE Trans. Patt. Anal. Mach. Intell. **18** (1996), no. 7, 673–689.

[8] A. H. Kam and W. J. Fitzgerald, *A general method for unsupervised segmentation of images using a multiscale approach*, Proc. 6^{th} Euro. Conf. Comp. Vis. (Dublin, Ireland), June 2000, pp. 69–84.

[9] S. Karunasekera and N. Kingsbury, *A distortion measure for blocking artifacts in images based on human visual sensitivity*, IEEE Trans. Im. Proc. **4** (1995), no. 6, 713–724.

[10] N. G. Kingsbury, *Complex wavelets for shift invariant analysis and filtering of signals*, J. Appl. Comput. Harm. Anal. **10** (2001), no. 3, 234–253.

[11] C. Kotropoulos, E. Augé, and I. Pitas, *Two-layer learning vector quantizer for color image segmentation*, Signal Processing IV: Theories and Applications (J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, eds.), Elsevier, 1992, pp. 1177–1180.

[12] D. Martin, C. Fowlkes, D. Tal, and J. Malik, *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*, Proc. 8^{th} IEEE Int'l Conf. Comp. Vis. (Vancouver, Canada), July 2001, pp. 416–425.

[13] C. Meilhac and C. Nastar, *Relevance feedback and category search in image databases*, IEEE Int'l Conf. Multimedia Computing and Systems (Florence, Italy), vol. 1, June 1999, p. 9512.

[14] D. Melas and S. Wilson, *Double markov random fields and bayesian image segmentation*, IEEE Trans. Sig. Proc. **50** (2002), no. 2, 357–365.

[15] E. Müller, W. Müller, S. Marchand-Maillet, D. Squire, and T. Pun, *A web-based evaluation system for content-based image retrieval*, Proc. ACM Multimedia Workshop on Multimedia Information Retrieval (Ottawa, Canada), October 2001, pp. 50–54.

[16] H. Müller, W. Müller, D. Squire, S. Marchand-Maillet, and T. Pun, *Performance evaluation in content-based image retrieval: Overview and proposals*, Patt. Rec. Lett. (Special Issue on Image and Video Indexing) **22** (2001), no. 5, 593–601.

[17] T. Pappas, *An adaptive clustering algorithm for image segmentation*, IEEE Trans. Sig. Proc. **40** (1992), 901–914.

[18] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, *Relevance feedback: A power tool for interactive content-based image retrieval*, IEEE Trans. Circuits and Video Tech. **8** (1998), no. 5, 644–655.

[19] Y. Rui, T. Huang, and S-F. Chang, *Image retrieval: Current techniques, promising directions and open issues*, Journal of Visual Communication and Image Representation **10** (1999), no. 4, 39–62.

[20] C. W. Shaffrey, N. G. Kingsbury, and I. H. Jermyn, *Unsupervised image segmentation via markov trees and complex wavelets*, Proc. Int'l Conf. Im. Proc. (Rochester, U.S.A.), September 2002.

[21] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, *Content-based image retrieval at the end of the early years*, IEEE Trans. Patt. Anal. Mach. Intell. **22** (2000), no. 12, 1349–1380.

[22] J. R. Smith and C.-S. Li, *Image retrieval evaluation*, IEEE CVPR98 Workshop on Content-Based Access of Image and Video Libraries (Santa Barbara, USA), June 1998, p. 112.

[23] D. Squire, H. Müller, W. Müller, S. Marchand-Maillet, and T. Pun, *Design and evaluation of a content-based image retrieval system*, Design and Management of Multimedia Information Systems: Opportunities and Challenges (S. M. Rahman, ed.), Idea Group Publishing, PA, USA, 2001, pp. 125–151.

[24] L. Yang, F. Albregtsen, T. Lønnestad, and P. Grøttum, *A supervised approach to the evaluation of image segmentation methods*, Int'l Conf. Comp. Anal. Images and Patterns (Prague, Czech Republic), September 1995.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399