Formal Semantic Models for Images and Image Understanding

Duc Do Audrey Tam
School of Computer Science and Information Technology, RMIT University
GPO Box 2476V
Melbourne, Australia 3001
{ducdo,amt}@cs.rmit.edu.au

ABSTRACT

A number of formal models for images [13,27,28] and models for text and image matching [1] have been proposed, but they have not sufficiently dealt with features with high-level semantics. While formal models are supposed to be precise, their structures should allow for the level of subjectivity involved in interpreting the high-level semantics inherent in images.

In our earlier work, we have shown that by restricting image retrieval to a specific domain, we can use logical reasoning based on common sense knowledge bases and the knowledge extracted from text corpora from the same domain to infer higher level semantics from lower level semantics. The interpretation of these lower level semantics, usually involving objects in the image, is subject to a lower level of subjectivity, hence making it possible to build an image model that is reasonably *objective*.

Based on these observations, we propose that an effective and feasible approach to build high-level semantics into image retrieval is to build semantic models for both the image (the object of meaning) and image understanding (the perception of meaning). The image model will aim to capture image features which are commonly accepted within a certain domain. The image understanding model will include mechanisms for subjective interpretation and will be associated with correspondence functions which measure similarity between instances of these two models. This level of similarity, or the semantic distance, can be called the *semiotic gap*. Using this framework, the image retrieval problem can be deemed equivalent to the problem of defining a correspondence function that delivers the theoretically, or empirically, narrowest semiotic gap.

We propose to construct the formal image model based on the concepts of semiotic structures, and an image understanding model based upon insights into how knowledge inference could assist with image retrieval. In this paper, we present the formal image model and argue why this model is suitable for the retrieval of visual data. An image understanding model, which is under ongoing research, is also briefly discussed with results of some preliminary experiments.

Keywords

Formal image model, semantic model, image understanding, logical reasoning, knowledge base.

First published at COSIGN 2004
14 – 16 September 2004, University of Split (Croatia)

1. INTRODUCTION

Popular image retrieval approaches are either content-based (CBIR) or text-based. Neither approach addresses image semantics very well. CBIR's low-level features, such as colours, textures and shapes, fail to capture the high level semantics inherent in images. Text-based approaches, with retrieval strategies largely borrowed *as-is* from the text retrieval discipline, rely on textual annotations to capture the meanings of images. Proposed methods using either approach have so far not been able to overcome the semantic gap problem [31]. As the semantic gap problem is fundamental to the problem of not being to satisfy a certain information need, the general performance of state-of-the-art image retrieval is very poor compared to what can be currently achieved in text retrieval.

Most work in CBIR has taken the view that low-level features are objective. This however is not entirely true because a photo image may not necessarily represent the *true nature* of the physical environment captured in its visual frame (although philosophically speaking, whether humans ever attain an understanding of the true nature of anything is questionable). For example, the colour of someone's hair in a photo may appear a lot lighter if the photo is well lit. The dominant colour feature for this person's hair expressed in some RGB value may not result in an accurate representation of the true nature. In some situations, a photographer may choose to focus on a particular object and leave out significant parts of the surroundings, resulting in a photo that misleads viewers' interpretation of the image content. Therefore a photo image is, at best, only a subjective representation of the true nature. The level of subjectivity is dependent on the equipment being used, the technical conditions of photography, such as the source of lighting or the film development technique, and the photographer's intention. Even outside the photographic domain, human visual perception already modifies true nature, as the human brain "has no direct access to information about properties of the external environment; it has access only to whatever information is available in the retinal image" [4]. Rossotti 1983 [29] illustrated this point by describing how mackerels appear silvery-green while in fact their scales are colourless, or how rainbow trouts appear to change colours when viewed from different angles. This is not even taking into account certain visual deficiencies such as colour-blindness, true to Kant's philosophy that "all our knowledge begins with the senses, proceeds then to the understanding, and ends with reason" [20].

Text-based approaches rely on language forms for both classification and retrieval. Although the interpretation of these forms can be objective through spelling and grammatical rules, the interpretation of their meanings can be subjective.

Taking into account these observations, we can say that every single visual feature in images is subjective. Therefore, an effective image retrieval model must be able to differentiate between and work with both semantic features and possible representations of true nature. A formal model with mechanisms to store, compare and retrieve features is not sufficient. It must be able to associate features with meanings in some forms of semiotic structures, analyse the entire feature space as a whole and reason with them in order to infer other meanings. This has given rise to the need of a formal semantic model for the image. As the baseline accuracy of CBIR is generally limited and often worse than that of keyword-based methods [23], this semantic model needs to be built on the foundation of textual annotations with provisional support for CBIR features.

2. FORMAL SEMANTIC MODEL FOR THE IMAGE

2.1 Prerequisites

An effective formal semantic model for the image should satisfy the following requirements:

- *It should support all image features, from primitive to abstract.* For the model to be complete, primitive features need to be supported as under certain conditions, they have been proven to be effective for certain search problems.
- *It should map naturally to the human search strategy*. The reason for this is obvious, as the well-researched semantic gap problem [14,31] is caused by two main deficiencies: the annotation deficiency and the query language deficiency. It is unnatural to expect humans to think in terms of low-level features. Squire and Pun 1998 [32] has demonstrated that algorithms which compute the similarity between images often make judgments that fail to match those of the human researcher, unless the human has been well trained in the computer algorithm.
- *It should be able to capture domain-specific high-level semantics*. Cyc is an example of a knowledge base system that helps constrain a search by tying knowledge to different domains, or microtheories [22]. The growth in both quantity and quality of Cyc has partly validated this approach. Technically, microtheories are only namespaces. However, semiotically, microtheories allow disambiguation of concepts given domain constraints. As image collections can be viewed as a knowledge base, they should also inherit this property.
- *It should support the incremental addition and modification of features as the domain grows (either in quantity or in quality)*. Compilers of dictionaries, year in year out, have to deal with the problem of introducing new words or expressions into dictionaries, or taking words out (albeit less often than the former activity) of them. It is the same problem when dealing with a body of knowledge related to a specific domain. Sometimes these domain shifts can be quite overarching, such as the departure of mathematics from the field of philosophy in modern times. In the old days, philosophers such as René Descartes or Alfred North Whitehead were also philosophers, making the study of mathematics part of the study of philosophy.
- *It should support unlimited levels of abstraction granularity*. A typical annotation structure, such as one found in the Lonely Planet image collection [6], that relies on metadata has a fixed number of levels for knowledge categorisation. This is partly implemented as an effort to compromise between information need and annotation efforts. However, the domain in question constantly changes, and as the body of knowledge may unpredictably expand in both breadth and depth, a fixed number of category levels is inadequate.

2.2 Success Criteria

The proposed image model should not be aimed at solving the general information retrieval problem, such as one offered by the Google text search engine [12]. Attempting to do this would fall into the *context dilemma* trap [7]. At the same time, if the theoretical model has any fixed requirements that effectively prevent it from practical adoption, such as requiring an excessive amount of time to annotate each image, it should not be recommended either.

Other success criteria of the model implementation should include:

- *It must be testable with real photo images as opposed to synthetic images*. Low-level features in synthetic images such as colours and shapes are usually well grouped, hence more easily identified, potentially resulting in a positive bias for content-based methods. We assume that a model that works well with real photo images will also work well with synthetic images.
- *Improvement in image retrieval using the model must be measurable*. The reason for this is obvious. These measures may include traditional information retrieval parameters such as recall and precision. Based on our own observations from the usage of the Lonely Planet image collection [24], the general aim is to attain significant and consistent improvement in recall, and *ideally* higher precision, for all image queries.
- *Improvement in image retrieval using the model must be comparable to other methods*. This criterion will depend on the availability of a common image benchmarking framework. The image retrieval community still does not have anything close to the TREC tracks [34] that are widely accepted and used in text retrieval experimentation and benchmarking. The leading image benchmarking framework, the Benchathlon Network [26], is still at its early stage and currently only supports CBIR methods. However, the objectives of the Benchathlon project are consistent with what we need for a suitable benchmarking framework. These include a standard data collection, a set of standard queries, a form of ground truth, a benchmarking engine, a set of performance measures and a standard access protocol.

2.3 Model Building Blocks

In text retrieval, everything can be traced back to terms and documents. A text collection consists of documents and a document consists of terms. At least that is the current basis for most, if not all, text retrieval methods. The content, information quality, and information quantity of a document can be examined by analysing these terms. Despite the issues with word sense and syntactical ambiguities, these terms,

belonging to specific subsets of a language or languages, already render themselves semantically interpretable.

There is no such simple equivalence in image collections, even though from a semiological point of view, images can be viewed as part of a language of signs [2]. If each image is viewed as a document, what forms of *visual units* can be deemed equivalent to terms in text? As we are dealing with two completely different forms of media, such equivalence may never be found. Some traditional image annotation approaches that try to associate a whole image with phrases or sentences (the simple captioning approach) are problematic. Frege, according to Haaparanta 1985 [17], observed that neither a word nor a noun phrase nor a verb phrase in isolation does tell us something. Wollheim 1996 [35] suggested that assimilating pictorial art, which can be quite unstructured, with textual annotations in forms of sentences, which are always structured, is not always possible. Because of these observations, many authors have tried breaking down the structure of the image by proposing their ideas of visual units, each aiming to solve problems related to their own domain. In pattern recognition, a field closely related to image retrieval, visual units are regarded as assemblies of low-level physical features such as pixels (with values denoting colours) or lines or shapes. Although these visual units, derived from formal analysis, are part of the basic vocabulary used in most art classes, they cannot be used in themselves to describe artistic or esthetic qualities in a visual piece of work. Bloomer 1990 was quite emphatic in proving this point in stating that "in its simplest terms, saying that all art consists fundamentally of line, shape, and color is like saying that all food is made up of protein, carbohydrates, and fats" [3].

To define visual units that can be used in a semantic model, we work on a basic assumption that domain experts have to be able to think in terms of these visual units. In fact, domain experts should not have to translate their information needs into other forms to be able to perform retrieval. The most straightforward way to achieve this is to define visual units that capture units of knowledge that are relevant to a particular domain. Johnson 1987 [19] proposed a concept called *image schema*, which is a derivation of Kant's concept of *schema*. In Johnson's definition, an image schema is "a mental pattern that recurrently provides structured understanding of various experiences, [and] is available for use in metaphor as a source domain to provide an understanding of yet other experiences". The keywords and phrases to note in this definition are: *pattern, structured understanding, understanding of various experiences*, and *metaphor*. From these, we can infer that an image schema must:

- be reusable (being a pattern);
- have a structure and provide a structured process for interpretation;
- allow the aggregation of various pieces of knowledge, thus requiring some form of nesting structure;
- facilitate the mental process of comparison and instantiation (with and to other experiences, either relating to or symbolised by the same image schema or not).

With these explicit properties, we propose to use image schemata as visual units for the formal semantic model. Each schema must represent a concept that is significant to a particular domain. The same idea is being used in the Cyc knowledge base, where, for example, a constant called #$VisualImage assumes different meanings in different microtheories (MT's). In the #$BaseKB MT, it is a type of #$InformationBearingWavePropagation, each instance of which is an event in which visible light is generated in a particular pattern, which (does or might) contain information for an observer; while in the #$UniversalVocabularyMt MT, it is simply classified as a #$TemporalStuffType, which interestingly enough, focuses on the *temporality* facet of this concept. Linguistically, an image schema can also be classified as a kind of source domain [21], which provides a reference point for other concepts.

2.4 Terminology

In discussion of the semantic model, the following terms are used:

- *Image collection*: a collection specific to a particular domain. All images in a collection must be relevant to the domain in question.
- *Image*: an image in a collection.
- *Image schema* (or *schema*): the abstraction of a visual unit in an image. An image contains instances of image schemata. An image schema may contain other schemas. A schema may inherit properties from higher-level schemata. Schemata can thus be grouped and categorised into a tree structure.
- *Image schema instance* (or *schema instance*): an instantiation of a schema within an image. All instances are unique to the image they belong to. For example, we all have a certain mental picture of an "abstract dog" (not any specific dog). The instance of Boo the dog in Picture A is different from the instance of Rover the dog in Picture B. The instance of Boo the dog in Picture A is also different from the instance of Boo the dog (albeit the same dog) in Picture C.
- *Semantic features (*or *features)*: semantic properties of an image, a schema, or a schema instance.
- *Metadata*: properties of an image, not related to or derived from any schemata or instances.
- *Semantic relationships* (or *relationships*): include *schematic relationships* and *instantial relationships*. Schematic relationships denote relationships among image schemata, while instantial relationships denote relationships among schema instances. In the current model, we propose two types of relationships: *spatial relationships* and *interaction relationships*. Instantial relationships do not have to be instances of schematic relationships, as opposed to schema instances, which are always instantiated from some image schemata.

2.5 Semantic Features

Eakins and Graham 1999 [9] proposed that image queries be categorised into three types based on the information need. These same query categories can be extended for image features to make them applicable to annotations. While only three query types were proposed, we propose the following four types of features. Features on all four levels can be present in images, image schemata, and schema instances.

2.5.1 Level 1 Features – Primitive

Level 1 features correspond directly with the primitive features of CBIR, including colours, textures and shapes. These features are automatically indexed and included as part of image annotations.

2.5.2 Level 2 Features – Derived

Level 2 features are those that can be automatically computed from level 1 features. A number of authors have proposed that this is possible. Eidenberger and Breiteneder 2002 [10] conducted an experiment to show that some human-world features such as symmetry, geometry and harmony can be computed from a set of primitive features including edge histogram, colour histogram and dominant colour. However, as the data set being used in this experiment only contains synthetic images, it is hard to guess if similar results can be expected with a data set containing real photos. Zhao and Grosky 2000 [37] even proposed that latent semantic analysis (LSA), a popular text retrieval technique, can be used to transform low-level features to a higher level of meaning. In a collection where certain objects can be consistently expressed with a common set of low-level features, it is easy to see how this approach could be feasible. However, if we are dealing with a collection diverse in topics and visual representations, this technique may not apply very well. As the proposed semantic image model is designed to work with all collections, it should have a built-in mechanism to capture these derived features, instead of just relying on automatic computation.

2.5.3 Level 3 Features – Topical

Level 3 features contain visible objects that humans recognise in the real world. The majority of work on semantics in image retrieval has focused on these objects. In our model, these features correlate directly to the image schemata. It is important to note that, within a certain domain, the set of topical features that are of interest can be finite, thus allowing the definition of a comprehensive list of objects applicable to a certain domain. An example of this definition can be found in the travel domain, as suggested by I'Anson 2000 [18]. This observation implies that we are not trying to define a generic list of objects that apply to all images under all circumstances.

2.5.4 Level 4 Features – Abstract Derived

Level 4 features are derived from features from the lower three levels or from other level-4 features themselves. These features usually represent abstract concepts such as feelings, emotions or interpretations, and thus, are highly subjective. Being subjective implies an inference process; therefore these features are classified as *abstract derived* instead of being simply *abstract*. Some common level-4 features can be derived within a particular domain with a low level of subjectivity. For example, if nudity is considered offensive in a certain domain, a specific rule can be defined that links the *is-nude* property of the *person* schema (a topical feature) to *offensiveness* (an abstract derived feature). In another domain, *offensiveness* may be linked to *violence* (another abstract derived feature). In this case, the *violence* feature is derived from a combination of other features present in the image.

2.6 Semantic Relationships

Queries involving relationships among certain objects potentially present in images are popular in image retrieval. For example, one may want to locate images with groups of people *in front of* a certain type of building, or images with a child *riding* a horse. Therefore, an effective semantic model for the image ideally should have a built-in mechanism to capture these semantic relationships. In the current model, we propose two types of relationships: spatial and interaction. In our future work, we may propose the addition of other types of relationships. It is important to note the role of schematic relationships in particular. Relating image schemata together, schematic relationships represent mental patterns built on individual experience and cultural or educational background. These include, for example, the notions of "dog *chasing* cat" or "big fish *eating* small fish". We assume that some kinds of relationships relevant to a particular domain are stereotypical, and thus, can be captured by schematic relationships.

2.6.1 Spatial Relationships

Spatial relationships among objects are usually the first types of relationships to be considered. This is probably due to the fact that these relationships are in themselves visually related. Relating back to our proposal to use image schemata as visual units and considering their roles in the real world as an aid for physical navigation [19], it only seems natural that spatial relationships among these schemata must also be considered to make navigation more effective. As the definition of these relationships is dependent on the domain in question, spatial relationships need to be either precise or imprecise. Certain domains, such as architectural design, require certain kinds of images, such as drawings, to contain precise spatial relationships, such as the exact metric distance between two objects. Users of other domains are simply not interested in precise spatial relationships. In these domains, the conversion of a certain spatial relationship, such as *is-far-from*, into some metric value will depend on what constitutes common sense in that domain. *is-far-from* in circuit design can probably be thought of in terms of centimetres, while it is more logically expressed in terms of dozens of metres in interior design. A model of spatial relationships such as one proposed by Güesgen 1998 [16] may be appropriate for use in this image model.

2.6.2 Interaction Relationships

Another common form of relationships in image queries represents interaction among certain objects in images. For example, one may be interested in locating images of two men fighting each other, or images of an African performer playing a bongo. As these kinds of relationships may exist among more than two objects, some form of syntactical constraints need to be introduced to avoid the kind of syntactical ambiguity found in sentences such as *Monkeys beat up cows with roses*. In the current model, we propose a form of syntactical constraint based on a structured description proposed by Tam and Leung 2001 [33]. All interaction relationships can be expressed as 4-tuples of Agent-Action-Recipient-Object.

Table 1. Formal rules of the semantic model for the image

Element	Rule
x	y | *instance*(s,i) | *instance*(s) | *instance*(i)
y	i | s
id(x)	UNIQUE-ID
S	∑s
I	∑i
i	∑*metadata*(i) ∑*feature*(i) ∑*instance*(s,i) ∑*rel*(i)
metadata(i)	*metadata_time*(i) | *metadata_location*(i) | *metadata_technical*(i) | *metadata_origin*(i)
metadata_time(i)	(`metadata-time`[:METADATA-SUBCATEGORY] $TERM_{TEMPORAL}$*)
metadata_location(i)	(`metadata-location`[:METADATA-SUBCATEGORY] $TERM_{GEOGRAPHIC}$)
metadata_technical(i)	(`metadata-technical`[:METADATA-SUBCATEGORY] $TERM_{TECHNICAL}$)
metadata_origin(i)	(`metadata-origin`[:METADATA-SUBCATEGORY] $TERM_{ORIGIN}$)
feature(x)	*feature1*(x) | *feature2*(x) | *feature3*(x) | *feature4*(x)
feature1(x)	(`feature-1:`PRIMITIVE-PROPERTY $TERM_{PRIMITIVE}$ | NUMERIC)
feature2(x)	(`feature-2:`DERIVED-PROPERTY $TERM_{DERIVED}$ | NUMERIC)
feature3(x)	(`feature-3` $TERM_{TOPICAL}$ | *id*(*instance*(,)))
feature4(x)	(`feature-4` $TERM_{ABSTRACT}$)
instance(s,i)	∑*feature*(*instance*(s,i))
rel(i)	*spatial-rel*(i) | *interaction-rel*(i)
spatial-rel(i)	(`rel-spatial`:$TERM_{SPATIAL}$ *id*($instance_a$(s,i)) *id*($instance_b$(s,i)))
interaction-rel(i)	(`rel-interaction`:$TERM_{INTERACTION}$ *id*($instance_a$(s,i)) *id*($instance_b$(s,i)) [*id*($instance_c$(s,i))])

* $TERM_O$ denotes a term taken from domain ontology O. For example, $TERM_{GEOGRAPHIC}$ is a term from a geographical ontology.

- An *Agent*, which is a schema instance in the semantic role of a person or thing that is the doer of an event, for example, ***girl*** *running*. An Agent is compulsory in an interaction relationship.
- An *Action*, which is a verb in the present continuous tense, indicating what the Agent is doing, for example, *running* or *sleeping*. An Action is compulsory in an interaction relationship.
- A *Recipient*, which is a schema instance, indicates the recipient (indirect object) of an action, for example, *boy giving* ***girl*** *guitar*. A Recipient is optional in an interaction relationship, such as in *child sleeping*.
- An *Object*, which is a schema instance, indicates the direct object of an action, for example, *girl holding* ***doll***. An Object is optional in an interaction relationship, such as in *Mark bullying John*.

This simple structure supports the definition of interactions among all possible schema instances in an image.

2.7 Metadata

Metadata are not exclusive to images. They are present in most forms of information media. Therefore the issues related to metadata should not be solved within the image retrieval discipline. However, as they are part of the proposed model, they are briefly discussed here for completeness purposes.

Units of information are often tagged with metadata to facilitate classification and retrieval. With images, there are certain properties that are not dependent on the content of the image itself but need to be captured as they are important to a particular domain. These may include the name of the photographer, the type of film the photo was captured on, the details of the camera equipment, the film development technique, or the geographic location where the photo was shot. In some cases, certain features can be inferred from certain metadata or vice versa, such as a certain photographer's name (a metadata field) and his topic of photography. The same situation exists between a landmark building (a topical feature) and a geographic location (a metadata field). However, it would be unnecessary to include these particular kinds of relationship explicitly, as they can be inferred from examining all the images within a certain collection. For example if all images with the *photographer* metadata of value Photographer A include a *war* feature, we can establish a relationship between Photographer A and the *war* topic in photography.

2.8 Formal Semantic Model

We propose the following formal semantic model for the image.

Let T be the collection of all textual terms used in image notations, and t_k T be a term in that collection, k {1, ... , u} where u = *size*(T).

Let D be a particular domain relevant to the image collection.

Given I as a domain-dependent image collection, i_l I is an image in that collection, and l {1, ... , w}, and w = *size*(I). As all images must be relevant to domain D, the relationship between images and the domain can be expressed through the following semiotic function:

truth(i | D) ≥ _, where _ is the truth threshold of domain D, which determines the minimum relevance value of any piece of information in the domain. A normalised value of function *truth*(i | D) = 1 signifies that image i is relevant to all kinds of information need within domain D, i.e. it should be returned for all queries. As this case never happens, 1 is only a theoretical value.

Given S as the collection of all schemata relevant to domain D, s S is a schema in that collection. As all schemata must be relevant to domain D, the relationship between schemata and the domain can be expressed through the following semiotic function:

truth(s | D) 8 _. A normalised value of *truth*(s | D) = 1 signifies that the image signifies that schema s is relevant to all queries of type *instance*(s) (there exists at least an instance of schema s) within domain D.

A schema instance can be expressed by the instantiation function *instance*(s). As ∑ s S are relevant to domain D, the collection of all instances is also relevant to domain D. This collection can be expressed as follows:

∑ $instance_p$(s) . ∑ s_q, where:

p P and P {1, ... , m}, and

q Q and Q {1, ... , n}, and

m = *size*(S), and

n = *max*(∑ *size*(*instance*(s_q))).

Metadata can be represented by the following metadata function:

$metadata_j$(i) = $_k$, where:

j J and J {1, ... , v}, and

v = ∑ $metadata_j$(i) x ∑ i_l

Level-1 features can be represented by the following feature function:

feature1(x) = $_k$, where x can be i, s or *instance*(s).

Features on other levels can be similarly defined by the functions:

feature2(x), *feature3*(x), and *feature4*(x).

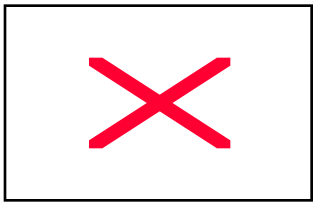
Relationships can be represented by relationship functions of the form *rel*(y_1, y_2, y_3) where y_1, y_2, and y_3 can be either s or *instance*(s).

Thus, an image can be seen as an aggregate of its metadata, features, instances and relationships:

i = ∑*metadata*(i) ∑*feature*(i) ∑*instance*(s,i) ∑*rel*(i)

Table 1 summarises the formal rules of this model.

Figure 1. A formal semantic representation



```
(image
        (metadata-origin:photographer “Richard Mills”)
        (metadata-technical:camera “Pentax 6x7”)
        (metadata-location:state-province Britanny)
        (metadata-location:country France)
        (feature-1:dominant-colour 0x32FE11)
        (feature-2:orientation horizontal)
        (feature-2:light-direction right)
        (feature-3 #ID1001)
        (feature-3 #ID1002)
        (feature-3 #ID1003)
        (rel-spatial:near #ID1001 #ID1002)
        (rel-spatial:right-of #ID1001 #ID1003)
        (rel-interaction:riding #ID1001 #ID1002)
        (rel-interaction:riding #ID1003 #ID1004)
        (rel-interaction:riding #ID1005 #ID1006)
)
(schema
        (id #SID1001)
        (feature-3 person)
)
(instance
        (id #ID1001)
        (instance-of #SID1001)
        (feature-3 gender-male)
)
(instance
        (id #ID1003)
        (instance-of #SID1001)
        (feature-3 gender-female)
)
```

2.9 Example of Model

There are different ways to relate all the descriptors together in a single model of the image. These include extensible markup languages such as XML, or description logics. We do not believe that a particular representation language is more suitable than others, but as logical reasoning will form part of the image retrieval algorithm based on this model, we use a form of description logics for descriptors to make them more suitable to perform inference functions with in the future.

Figure 1 contains an example which illustrates the model presented in the previous section.

3. IMAGE RETRIEVAL PROCESS AND FORMAL SEMANTIC MODEL FOR IMAGE UNDERSTANDING

This section discusses work in progress and is included here for completeness purposes. It aims to put the model proposed in the previous section into the overall context and explains how it can be used in the retrieval process.

3.1 Information Retrieval as an Interactive Process

Conventional information retrieval paradigms assume that it is the searcher's responsibility to properly communicate the information need to the information retrieval system. This assumption has led to construction of retrieval systems as depicted in Figure 2, in which the smilie symbol represents the searcher and the computer symbol represents the retrieval system.

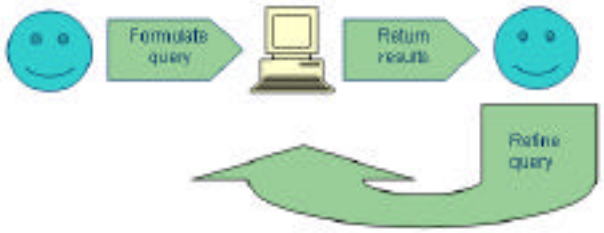


Figure 2. Conventional image retrieval paradigm

The key feature to note in these systems is that the system assumes that the query already contains sufficient information for search to take place. There are issues even with relevance feedback mechanisms where the searcher is asked to refine the choices after being presented with the first set of results. Searchers are usually asked to refine results when their queries result in too many matches. Most users may be, in fact, reluctant to do this because they are already overwhelmed by the amount of information returned by the first search attempt. Another issue is that if the first set of results being returned is deemed of little relevance to the user's information need, it may be perceived as a negative experience for the user and may, thus, dissuade the user from further interaction with the system.

It is quite obvious when analysing information retrieval scenarios between humans that the above model is rather unnatural and awkward. Let's consider a non-electronic environment, a non-electronic library, where patrons call upon the assistance of reference librarians to locate a certain publication on a certain topic. Slavens [30] edited a collection of reference interviews and questions that were manually collected in real libraries. Scrutinising this collection reveals a range of information needs and different methods that librarians used to fulfill an information need. Table 2 outlines a reference interview between a librarian (L) and a patron (P). These roles can be mapped to a typical image retrieval scenario where the user P issues queries and the system L responds with matches deemed relevant to these queries.

From the above reference interview, we have reached the following observations:

- Queries contain certain assumptions that are not necessarily correct. A good retrieval strategy should allow for the detection of assumptions and a method to clarify them if possible.

Table 2. Semantic analysis of a reference interview

Question/Answer in Sequential Order	Analysis of Information Flow
P: Do you have a history encyclopaedia?	P makes an assumption that the topic is historically related. Queries can contain assumptions that are not necessarily correct
L: What was it that you had in mind to look up?	Good strategy to clarify assumptions.
P: I need a chapter on the Gold Rush.	Query contains semantic ambiguity – at least on geographical and temporal levels. "Chapter" may tell the system something about the amount of information required.
L: In the United States?	L forms an appropriate question to help clarify semantic ambiguity related to geographic location.
P: Yes.	Query *clarity* is improved through geographic disambiguation.
L: Do you know when it happened? There were several I think.	L forms an appropriate question to help clarify semantic ambiguity related to time period.
P: The one I want was around 1848.	Query *clarity* is improved again through temporal disambiguation.
L: Is there any other word we could use for Gold Rush?	L tries to extract more information in attempt to improve query *quality*.
P: Sometimes they're called the Forty-Niners. Where would you find this information?	P is able to provide a crucial piece of information in "the Forty-Niners" which may help retrieval precision. The last question, though, may signal a certain level of impatience from P, probably due to being asked too many questions.

- Queries can be semantically ambiguous. These ambiguities can be multi-fold. A good retrieval strategy should allow for the detection of semantic *clashes* and a method to resolve them as early in the process as possible.
- A good retrieval strategy should *encourage* users to build more quality into queries even before attempting search.
- A good retrieval strategy is a collaborative process where the system and the user both contribute to the task of locating the relevant information.

In attempting to engage the searcher in clarifying semantic ambiguities in query terms, the Getty Image Search system [11] reflects this more natural retrieval model to some extent.

Based on this observation, we propose that query analysis should be fundamental to effective information retrieval models in general, and image retrieval models in particular. In Figure 3, we propose a more natural paradigm for image retrieval.

In Figure 3, the "Ambiguous?" conditional can be determined by a function that measures ambiguity and checks if the

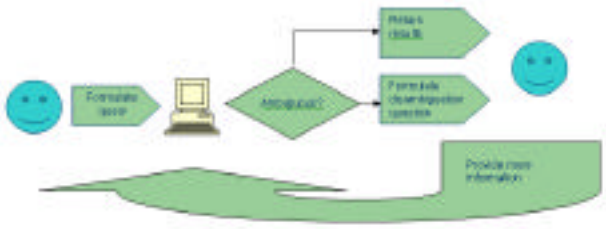


Figure 3. Interactive image retrieval paradigm

ambiguity in the current query is below a certain threshold. Relevant ideas may need to be extracted from the development in the field of query clarity [5] in order to develop this function.

3.2 Image Retrieval as a Reverse Function of the Visual Perception Process

In Figure 4, the features depicted in the right hand frame (perceived shapes and colours) map directly onto level-1

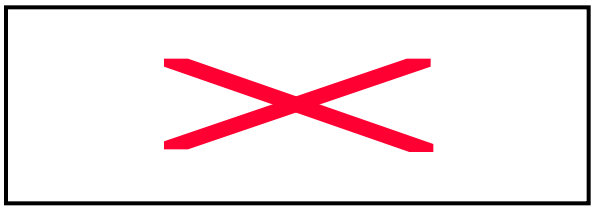


Figure 4. Perception of object shapes

features in the image model. Generalising on this observation, we can argue that the image retrieval process is a reverse function of the visual perception process.

Building an image retrieval system therefore implies building perception and reverse-perception functions for all levels of features in the image model. We call the collection of these functions the *formal semantic model for image understanding*. In order to build this model, we will need to examine how features can be computed from other level features. This is where work such as one proposed by Eidenberger and Breiteneder 2002 [10] (computing level-2 features from level-1 features) and one proposed by Do and Tam 2004 [7] (computing level-4 features from level-3 features) fit in. Through empirical analysis [7,8], we have demonstrated that domain-specific text corpora contain knowledge that can be used in improving image retrieval performance, and that logical reasoning using common sense and domain-specific knowledge bases may be used as a method to infer new features given existing features.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a formal semantic model of the image. For this model to be used effectively in image retrieval, a formal semantic model for image understanding also needs to be developed. Part of this model is the definition of a correspondence function which computes the semiotic gap between instances of the image model and the image understanding model. This particular model will be presented in another paper in the near future.

Related to the image model itself, we have only outlined the concepts and the formal rules for image representations. We have not considered the specification of specific ontologies that are suitable for use with the model. This consideration and, ideally, the recommendation of specific ontologies, if they exist, is part of the ongoing research. These ontologies will need to be structured around a common knowledge framework, such as one based on elementary typics of knowledge proposed by Gudwin and Gomide 1997 [15]. Johnson 1987 [19] classified image schemata into different types such as center-periphery, containment, part-whole, and verticality. This system of classification may be used to make schema instantiations more specific.

Upon completion of both models, experiments will be conducted with different image collections to validate our approach. The authors also expect that the existence of these two formal models may facilitate the creation of a formal image retrieval benchmarking framework for both CBIR and text-based techniques.

5. ACKNOWLEDGMENTS

The authors acknowledge the assistance of Dr. Ron Gallagher and Lonely Planet Pty. Ltd. in providing them with sample images and guidebook data that are used in the experiments referred to in this work and in other related works conducted by the authors.

All photographs used in this paper, unless otherwise specified, are obtained with permission from Lonely Planet Images.

6. REFERENCES

[1] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I., Matching words and pictures. In *Journal of Machine Learning Research 3*, pp.1107-1135, 2003.

[2] Barthes, R. *Elements of Semiology* (trans. Annette Lavers and Colin Smith). London: Jonathan Cape, 1967.

[3] Bloomer, C. B. *Principles of Visual Perception*. The Herbert Press, 2nd edition, p.211, 1990.

[4] Boothe, R. G. *Perception of the Visual Environment*. Springer-Verlag, New York, p.13, 2002.

[5] Cronen-Townsend, S., and Croft, B. Quantifying query ambiguity. In *Proceedings of the Conference on Human Language Technology (HLT)*, 2002.

[6] Do, D. *Building High-Level Semantics into Lonely Planet's Image Collections*. Industry Presentation, July 2004.

[7] Do, D., and Tam, A. M. The context dilemma and observations on using vocabulary constraints in image retrieval. *Paper submitted to CIKM*, 2004.

[8] Do, D., and Tam, A. M. Querying compound concepts in image collections. In *Proceedings of the International Conference on Intelligent Agents, Web Technology and Internet Commerce (IAWTIC)*, Gold Coast, Australia, July 2004.

[9] Eakins, J. P., and Graham, M. E. *Content-based Image Retrieval: A report to the JISC Technology Applications Programme*. Institute for Image Data Research, University of Northumbria at Newcastle, January 1999.

[10] Eidenberger, H., and Breiteneder, C. Semantic feature layers in content-based image retrieval: implementation of human-world features. In *Proc. ICARCV*, 2002.

[11] *Getty Images*. http://www.gettyimages.com. Last accessed April 2004.

[12] *Google*. http://www.google.com. Last accessed August 2004.

[13] Grosky, W. I., and Stanchev, P. L. An image data model. In *Advances of Visual Information Systems, Visual 2000, 4th International Conference, Lyon, France*, pp.14-25, 2000.

[14] Gudivada, V. N., and Raghavan, V. V. Content-based Image Retrieval Systems. In *IEEE Computer, volume 28, issue 9*, pp.18-22, September 1995.

[15] Gudwin, R. R., and Gomide, F. A. C. An approach to computational semiotics. In *Proceedings of the 1997 International Conference on Intelligent Systems and Semiotics*, NIST Special Publication 918, pp. 467-470, 1997.

[16] Güesgen, H. W. Spatial reasoning based on Allen's temporal logic. In *International Computer Science Institute Technical Report TR-89-049*, 1998.

[17] Haaparanta, L. Frege's context principle. In *Communication and Cognition 18*, p.81-94, 1985.

[18] I'Anson, R. *Travel Photography: A Guide to Taking Better Pictures*. Lonely Planet Publications, 1st edition, pp.100-204, October 2000.

[19] Johnson, M. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press, 1987.

[20] Kant, I. *et al*. *Critique of Pure Reason*. Hackett Publishing Company, December 1, 1996.

[21] Lakoff, G. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago, pp. 276-277, 1987.

[22] Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., and Shepherd, M. Cyc: Toward programs with common sense. In *Communications of ACM, vol. 33, no. 8*, August 1990.

[23] Li, M., Chen, Z., and Zhang, H.-J. Statistical correlation analysis in image retrieval. In *Pattern Recognition 35*, pp.2687-2693, 2002.

[24] *Lonely Planet Image Collection*. http://www.lonelyplanetimages.com. Last accessed June 2004.

[25] Loos, E. E., Anderson, S., Day., D. H. Jr., Jordan, P. C., and Wingate, J. D. (eds.). *Glossary of Linguistic Terms*. Extract from the LinguaLinks Library, SIL International, CD-ROM version 5.0, 2003.

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

[27] Mechkour, M. A multifacet formal image model for information retrieval. In *Proceedings of the Final Workshop on Multimedia Information Retrieval (Miro '95)*, Glasgow, Scotland, 18-20 September 1995.

[28] Meghini, C., Sebastiani, F., and Straccia, U. A model of multimedia information retrieval. In *Journal of the ACM (JACM), volume 18, issue 5*, pp.909-970, September 2001.

[29] Rossotti, H. *Colour*. Penguin Books, Great Britain, p.92, 1983.

[30] Slavens, T. P. (ed.). *Reference Interviews and Questions*. Michigan University: Campus Publishers, 3rd Ed., no dates noted for this edition.

[31] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-based image retrieval at the end of the early years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, December 2000.

[32] Squire, D. M., and Pun, T. Assessing agreement between human and machine clusterings of image databases. In *Pattern Recognition, vol. 31-32*, pp.1905-1919, 1998.

[33] Tam, A. M., and Leung, C. H. C. Structured natural-language description for semantic content retrieval. In *J. American Soc. Information Science*, pp. 930-937, September 2001.

[34] *TREC Tracks*. http://trec.nist.gov/tracks.html. Last accessed August 2004.

[35] Wollheim, R. On the assimilation of pictorial art to language. In *Towards a Theory of the Image* (ed. Jon Thompson), Jan van Eyck Akademie, Maastricht, p.27, 1996.

[36] Zhao, R., and Grosky, W. I. From features to semantics: some preliminary results. In *Proc. International Conference on Multimedia and Expo*, 2000.