Reasoning about Form and Content of Multimedia Objects (Extended Abstract)

Carlo Meghini, Fabrizio Sebastiani and Umberto Straccia*

1 Introduction

Due to the pervasive role of multimedia documents (MDs) in nowadays information systems, a vast amount of research has been carried out in the last few years on methods for effectively retrieving such documents from large repositories. This research is still in its infancy, due to the inherent difficulty of indexing documents pertaining to media other than text in a way that reflects their information content and, as a consequence, that significantly impacts on retrieval. Nonetheless, a number of theoretical results concerning sub-problems (e.g. the image retrieval problem) have been obtained and experimented with, and on top of these a first generation of retrieval systems have been built [10] and, in some cases, even turned into commercial products [2, 8].

The distinguishing feature of these multimedia retrieval systems (MRSs), and of the related research models, is the lack of a proper representation and use of the *content* of non-textual documents: only features pertaining to their *form*, being most amenable to automatic extraction through digital signal processing (DSP) techniques, are used upon retrieval. But this is disturbing, as documents, irrespective of the representation medium they employ, are to be regarded as *information carriers*, and as such are to be studied along two parallel dimensions, that of *form* (or *syntax*, or *symbol*) and that of *content* (or *semantics*, or *meaning*). Here, "form" is just a collective name for all those (medium-dependent) features of an information carrier that pertain to the representation and to the representation medium, while "content" is likewise a collective name for those (medium-independent) features that pertain to the slice of the real world being represented, which exists independently of the existence of a representation referring to it. The main thrust of this paper is that a data model for the retrieval of MDs (which we here take as consisting of multiple sub-documents each pertaining to possibly different media, rather than as just non-textual "atomic" documents) not only needs both dimensions to be taken into account, but also requires that *each of them be tackled by means of the tools most appropriate to it*, and that these sets of tools be integrated in a principled way in order to ensure transparent user access. Concerning the issue of tool appropriateness, we think that, inasmuch as the techniques from DSP (used e.g. in image and audio retrieval) are inadequate to reason about content, those from the field of knowledge representation are inadequate to deal with document form.

This study addresses the problem of injecting semantics into MD retrieval by presenting a data model for MDs where sub-documents may be either texts or images. The way this model enforces the interaction between these two media is illustrative of how other media might also be accounted for. Texts and images are represented at the content level as sets of properties of the real-world objects *being represented*; at this level, the representation is *medium-independent*, and a unique language for content representation is thus adopted. This data model is logic-based, in the sense that this latter language is based on a *description logic* (DL – see e.g. [3]).

Texts and images are also represented at the form level, as sets of physical features of the objects *representing* a slice of the world; at this level, the representation is *medium-dependent*, so

*Authors' address: Consiglio Nazionale delle Ricerche, Istituto di Elaborazione dell'Informazione, Via S. Maria 46, I-56126 Pisa, Italy, E-mail: ⟨*lastname*⟩@iei.pi.cnr.it

different document processing techniques are used to deal with sub-documents expressed in the different media. Although features pertaining to form are not represented explicitly in the DL, they impact on the DL-based reasoning through a mechanism of "procedural attachments". This implements the connection between (logical) reasoning about content and (non-logical) reasoning about form. From the point of view of the *semantics* of the query language, this latter connection is established by restricting the set of interpretations of the logical language to those that verify the constraints imposed at form level by the results of the text processing and DSP analysis. This mechanism for giving semantics to procedural attachments, thereby allowing to effectively merge logical and non-logical reasoning, has also been called the *method of concrete domains* [1].

Although the main task of our DL is reasoning about content, our DL-based query language is also endowed with the referential machinery to address the form dimension of text and images; linking the form and content of the same document is made possible by the sharing of the same DL symbols. The DL-based query language thus allows the expression of retrieval requests addressing both structural (form) and conceptual (content) similarity, and its underlying logic permits, among other things, to bring to bear domain knowledge (whose representation DLs are notoriously good at) in the retrieval process. The query language also includes facilities for fuzzy reasoning, in order to address the inherently quantitative nature of notions like "similarity" between text/images or between their features (word morphology, image colour, image shape, and the like). The model is extensible, in that the set of symbols representing similarity can be enriched at will, to account for different notions of similarity and methods for computing it.

The resulting retrieval capability thus extends that of current MRSs with the use of semantic information processing and reasoning about text/image content. So far, the only attempts in this direction had been based on textual annotations to non-textual documents ("captions": see e.g. [17]), in some cases supported by the use of thesauri to semantically connect the terms occurring in the text [11]; this means that text is seen as mere comment on the non-textual document, and not as an object of independent interest and therefore subject to retrieval *per se*. In our model text and images are both first-class citizens, and this clearly indicates how the extension to other media could be accomplished.

The paper is organised as follows. Section 2 deals with the "form" dimension of texts and images, defining the notions of *text layout* and *image layout*; these consists of the symbolic representation of form-related aspects of a text or image. Both notions are endowed with a *mereology*, i.e. a theory of parts, based on the notion of *text region* and *image region* as from digital geometry. In Section 3 we briefly introduce a fuzzy DL, discussing its use to represent document content and to "anchor" content representations to form representations. Document bases are defined in Section 4, while Section 5 introduce queries, categorising them with respect to the representation medium and to the dimension involved, and describing how the "procedural attachment" and the "concrete domains" methods provide a smooth integration of form- and content-based retrieval.

The full paper also discusses its computational complexity and the realization of a MRS supporting the model. Concerning this latter point, we only remark that this are well within reach of the current technology. In particular, we have developed a theorem prover for a significant extension of the DL we use here [14], based on a sound and complete Gentzen-style sequent calculus; this theorem prover is currently being prototyped for subsequent experimental evaluation.

2 Representing form

We here briefly hint to the basic notions of the form dimension of texts and images; these notions are formally defined and more fully argued in the full paper.

Given an alphabet Σ, a *text layout* is a pair $\langle n, w\rangle$ (abbreviated as w_n) where $n \in \mathbb{N}$ is the *length* of the text and $w : \{0, n-1\} \to Seq(\Sigma)$ is a total function assigning a word to each position in $[0, n-1]$. A *(grounded) text region* is a pair $\langle w_n, S\rangle$, where S is a sub-interval of the $[0, n-1]$ interval. By *extended text region* we will mean a set of (non necessarily contiguous) text regions from the same layout. By $Seq(\Sigma)^n$ we denote the set of all possible text layouts of length n. The *text universe* $\mathcal{T} = \cup_{i\in\mathbb{N}_+} Seq(\Sigma)^i$ is the union of all finite-length text layouts.

Given a set of colours $\mathcal{C}$, an *image layout* is a triple $i = \langle A^i, \pi^i, f^i \rangle$, where A^i, the *domain,* is a finite, aligned, rectangular region (see e.g. [15, Chapter 11]); π^i is a partition of A^i into non-empty connected regions $\{T_1, ..., T_n\}$, called *atomic regions;* f^i is a total function from π^i to $\mathcal{C}$, assigning a colour to each atomic region (and therefore called the *colour function*) such that no two neighbour atomic regions have the same colour. Informally, by *extended region* we will mean a set of contiguous regions, and by *extended colour function* f^i_e of a layout i we will mean the function that extends the colour function to extended regions (returning so-called *colour distributions*). In general, a region S is not bound to a particular layout; this binding is realized in the notion of *grounded region*, which we define as a pair $\langle i, S \rangle$, where $i = \langle A^i, \pi^i, f^i \rangle$ is a layout and $S \in \pi^i_e$.

3 Representing document contents

We take the content of a document (be it text, or image, or any combination of the two) to be a *scene*, i.e. the set of all states of affairs consistent with the information in the document. Informally, the content of a text will thus be the set of situations that support the sentences making up the text, and that of an image will thus be a set of situations indistinguishable from the visual point of view. For instance, the scene denoted by image i might be the one in which Giulia is hugging Francesco, and in which they both wear blue; this scene encompasses all visually equivalent situations of this kind, i.e. it is irrespective of when the action takes place, and of what Francesco and Giulia are thinking of. Symbolic representations of scenes are important if an MD base is to be accessed for retrieval based also on the properties of the individuals referred to within documents. The formalism we have chosen for representing and reasoning on image contents is a *Description Logic* (DL – see e.g. [3, 13]). DLs (aka "Terminological Logics") are contractions of first order logic, and have an "object-oriented" character that makes them especially suitable to reasoning about structured objects. They have been shown adequate for modelling a large class of facts relevant to information systems [4, 5], and a number of them also have decidedly better computational properties than FOL (decidability and, in some cases, also polynomiality; see e.g. [6]).

The specific DL that we adopt is $\mathcal{ALC}$ [16], a significant representative of the DLs family; however, our model is not tied in any way to this particular choice, and any other DL would easily fit in it. The language of $\mathcal{ALC}$ includes unary and binary predicate symbols, called *primitive concepts* (indicated by the metavariable A with optional subscripts) and *primitive roles* (metavariable R), respectively, and individuals (metavariable a). These are the basic constituents by means of which *concepts* (metavariable C), i.e. "non-primitive predicate symbols", are built via *concept constructors*, according to the following syntactic rule:

$$C \longrightarrow A \mid C_1 \sqcap C_2 \mid C_1 \sqcup C_2 \mid \neg C \mid \forall R.C \mid \exists R.C$$

A *crisp assertion* is an expression having one of the forms: a) $C(a)$, meaning that a is an instance of C; b) $R(a_1, a_2)$, meaning that a_1 is related to a_2 by means of R; c) $T \sqsubseteq T'$, where T and T' are both concepts or both roles, means that T is a specialization of T'. The first two kinds of assertions are called *simple assertions,* while any instance of the last kind is said to be an *axiom.*

In order to deal with the uncertainty inherent in similarity-based retrieval, we add to $\mathcal{ALC}$ *fuzzy assertions* (see e.g. [7]), i.e. expressions of the form $\langle \alpha, n \rangle$ where α is a crisp assertion and $n \in [0, 1]$, meaning that α is true "to degree n". We will use the terms *fuzzy simple assertion* and *fuzzy axiom*, with the obvious meaning. The semantics of fuzzy $\mathcal{ALC}$ is detailed in the full paper.

Let us now discuss how fuzzy $\mathcal{ALC}$ is used for document content representation. We recall that content is a medium-independent notion, and as such may be discussed without reference to whether the document analysed is an image or a text. Therefore, let i be a (text or image) layout uniquely identified by the individual `i`. A *content description* δ for i is a set of fuzzy assertions, consisting of the union of four component subsets:

1. the *document identification,* a set containing a single fuzzy assertion of the form $\langle \texttt{Ego(i)}, 1 \rangle$, whose role is to associate, along with the layout naming function n_l (see Section 4), a content

description with the layout it refers to. In particular, in what follows $\sigma(\mathtt{i})$ will denote the set of the (possibly many) content descriptions whose identification is `Ego(i);`

2. the *object anchoring,* a set of fuzzy assertions of the form $\langle \mathtt{Rep}(\mathtt{r},\mathtt{o}), n\rangle$, where `r` is an individual that uniquely identifies a grounded (text or image) region of i and `o` is an individual that identifies the object denoted by the region;

3. the *scene anchoring,* a set of fuzzy assertions of the form $\langle \mathtt{About}(\mathtt{i},\mathtt{o}), n\rangle$, where `i` and `o` are as above. By using these assertions, an indexer can state what the scene represented by the document is about;

4. the *scene description,* a set of fuzzy simple assertions (where neither the predicates `Ego`, `Rep` and `About`, nor identifiers pertaining to layout such as the `i`'s and `r`'s above, occur), describing important facts described by the document about the individuals identified by assertions of the previous two kinds.

While the task of components 1 to 3 is that of binding the form and content dimension of the same document, component 4 pertains to the content dimension only. Note that there may be more than one content description for the same document i; this is meant to reflect the fact that there may be multiple viewpoints under which a document may be considered. As an example, let us consider a photograph showing a singer, Mary, performing as Zerlina in Mozart's "Don Giovanni". Part of a plausible content description for this image, named `i`, could be (for simplicity, in this example we only use crisp assertions):

{`Ego(i), About(i,o), Rep(r,mary), DonGiovanni(o), Plays(mary,zerlina)`}

4 Document bases

We model a document base as a collection consisting of a text base and an image base, plus additional information on "document structures" (DSs). This latter aspect, that for reasons of space we relegate to the full paper, includes a formal definition of the notion of DS, which is meant to address the possibly complex, hierarchical structure that a MD may have. Typically, leaf nodes of document structures will be elements of either the text base or image base.

An *image base* is a 4-tuple $IB = \langle IL, \nu_I, \Sigma_{IC}, \Sigma_{ID}\rangle$ where: a) IL is a set of image layouts; b) ν_I is a naming function mapping image layouts and grounded image regions in IL into individuals, which therefore act as unique names for them; c) Σ_{IC} is the set of content descriptions associated to the layouts in IL; and d) Σ_{ID} is the domain knowledge for the images in IB. In a completely analogous way, a *text base* is a 4-tuple $TB = \langle TL, \nu_T, \Sigma_{TC}, \Sigma_{TD}\rangle$, with the obvious meaning.

5 A unified query language

A query posed to a document base can refer to the structure of the requested documents, and/or to features of their text components, and/or to features of their image components. For reasons of space we will here address only the last aspect, as features concerning the first two can be figured out from the discussion that follows (and are to be found in the full paper).

A query addressed to an image base can refer either to the form dimension, in which case we call it a *visual* query, or to the content dimension, in which case we call it a *conceptual* query. These two categories are exhaustive but not disjoint. Visual queries can be partitioned as follows:

1. *concrete visual queries:* these consist of images themselves that are submitted to the system as a way to indicate a request to retrieve "similar" images;

2. *abstract visual queries:* these are "abstractions" of layouts that address specific aspects of image similarity via artificially constructed image elements; they can be further categorised into: a) *colour queries*, i.e. colour distributions that are used to retrieve images with a

similar colour distribution; b) *shape queries*, i.e. specifications of one or more shapes (closed simple curves in the 2D space) and possibly of their spatial relationships, used to retrieve images in which the specified shapes occur as contours of significant objects, in the specified relationships; and c) other categories, such as spatial and texture queries [9], which we do not deal with here.

Concrete visual queries are processed in a global way, i.e. by matching a vector of global features extracted from the query image with each of the homologous vectors extracted from the images subject to retrieval. Abstract visual queries replicate the same phenomenon but at a different level of granularity: the visual entities relevant to retrieval (such as shape and colour) are represented via features extracted from the query, and are to be matched with their homologous features extracted from images subject to retrieval. There are a number of different methods for performing image matching, each based on a specific set of features and a specific way for combining them in order to obtain a significant similarity assessment. These methods are mostly application-dependent, in that their effectiveness is a function of the type of targeted images and, most importantly, of the goal of retrieval, which greatly affects the relevant similarity criteria. In addition, they are better expressed at the procedural level, if anything else for efficiency considerations. For all these reasons our model, unlike [12], does not provide the machinery for defining similarity functions, but views them as "black boxes" which have two objects of the proper type as input and produce a degree of similarity, *i.e.* a number in $[0, 1]$, as output. These black boxes are the semantics of *special predicate symbols* (SPSs) which model similarity at the DL level. In this way, a *procedural attachment* is established between the logical symbols that denote visual features and the algorithms that compute them.

In order to query layouts, the following SPSs are introduced:

- *symbols for global matching:* in general, there will be a set of such symbols, each capturing a specific similarity criterion. Since from the conceptual viewpoint these symbols form a uniform class, we just include one of them in our language, to be understood as the representative of the whole class. Any other symbol of the same sort can be added without altering the structure and philosophy of the language. So, for global matching we use the SPS
  - `SI(i,j)` (standing for `Similar Image`): assesses the similarity between two layouts `i` and `j`;
- *symbols for local matching:* these come in two sorts. First we have *selectors,* which are SPSs needed to select the entity to match from a layout:
  - `HAR(i,r)` (`Has Atomic Region`): a selector relating the image `i` to any of its grounded atomic regions `r`;
  - `HR(i,r)` (`Has Region`): a selector relating the image `i` to any of its grounded regions `r`;
  - `HC(r,c)` (`Has Colour`): a selector relating the grounded region `r` to its colour `c`;
  - `HS(r,s)` (`Has Shape`): a selector relating the grounded region `r` to its shape `s`.

  Second, we have symbols for local matching, assessing similarity between local entities. Similarly for what it has been done for global matching, we include in the language one symbol for each category of entities to be matched; so we have:
  - `SC(c,c')` (`Similar Colour`): returns the similarity between two colour distributions `c` and `c'`;
  - `SS(s,s')` (`Similar Shape`): returns the similarity between two shapes `s` and `s'`.

The semantics of the symbols introduced so far is fixed, and is detailed in the full paper.

We are now in the position of defining the *query language* of the model. In order to comply with the philosophy of the model, the query language must satisfy two basic requirements. First, it has to be a concept language of a DL, so that matching queries against images can be done in the logical framework defined so far; and, second, it must respect the semantics of the symbols for addressing images introduced in the previous section.

In our query language (again, described in detail in the full paper) a query is a combination, via the conjunction and disjunction constructors, of so-called *image-concepts*, each of which may have one of four forms (following the order of the syntax):

1. a global similarity match request;
2. a query on some content-related object described by *content-concept*, which is any $\mathcal{ALCO}$ concept built with the symbols used for scene descriptions;
3. a query on an atomic region, which is required to satisfy the property expressed by the embedded *region-concept*;
4. a query on an extended region. In this case, the embedded concept is the same as a *region-concept*, but it must include a `Rep` clause; this prevents the specification of queries involving arbitrary extended regions of an image, of which there are an exponential number.

A *region-concept* gives conditions on a region, and is built as an $\sqcap/\sqcup$-combination of three basic conditions: one concerns the colour of the region, which must be the same as, or similar to, a specified colour (*colour-concept*); another analogously concerns the shape of a region(*shape-concept*); the third involves the individual represented by a region, and is a content concept.

Let us reconsider the example introduced in Section 3. The images that are about Don Giovanni are retrieved by the query $\exists\texttt{About.DonGiovanni}$. Those showing the singer Mary are described by $\exists\texttt{HR.}\exists\texttt{Rep.}\{\texttt{mary}\}$. Turning to visual queries, the request to retrieve the images similar to a given one, named `this`, is expressed by $\exists\texttt{SI.}\{\texttt{this}\}$, and can be easily combined with any conceptual query, e.g. yielding $\exists\texttt{SI.}\{\texttt{this}\} \sqcup \exists\texttt{About.DonGiovanni}$, which would retrieve the images that are either similar to the given one or are about Don Giovanni. As far as local visual queries are concerned, the images in which there is a blue region whose contour has a shape similar to a given curve `s` are denoted by the query $\exists\texttt{HAR.}(\exists\texttt{HC.}\{\texttt{blue}\} \sqcap (\exists\texttt{HS.}\exists\texttt{SS.}\{\texttt{s}\}))$. Finally, the user interested in retrieving the images in which Mary plays Zerlina and wears a bluish dress, can use the query $\exists\texttt{HR.}\exists\texttt{Rep.}(\{\texttt{mary}\}\sqcap\exists\texttt{Plays.}\{\texttt{zerlina}\})\sqcap(\exists\texttt{HC.}\exists\texttt{SC.}\{\texttt{blue}\})$.

Our image data model is based on the idea that, in response to a query Q addressed to an image base $IB = \langle L, n_l, n_r, \Sigma_C, \Sigma_D\rangle$, the layout named `i` is attributed a degree of relevance n iff:

$$n = max_{\{\delta_j \in \sigma(\texttt{i})\}}\{n_j = Maxdeg(\delta_j \cup \Sigma_D, Q(\texttt{i}))\}$$

Let us consider an image base containing two image layouts named `i` and `j`, such that:

$$\{\langle \texttt{Ego(i)}, 1\rangle, \langle \texttt{About(i,o)}, 0.8\rangle, \langle \texttt{DonGiovanni(o)}, 1\rangle\}$$
$$\{\langle \texttt{Ego(j)}, 1\rangle, \langle \texttt{About(j,o)}, 0.7\rangle, \langle \texttt{WestSideStory(o)}, 1\rangle\}$$

are in Σ_I. Moreover, Σ_A contains the following axioms:

$$\langle \texttt{DonGiovanni} \sqsubseteq \texttt{EuropeanOpera}, 1\rangle$$
$$\langle \texttt{WestSideStory} \sqsubseteq \texttt{AmericanOpera}, 1\rangle$$
$$\langle \texttt{EuropeanOpera} \sqsubseteq \texttt{Opera} \sqcap (\exists\texttt{ConductedBy.European}), 0.9\rangle$$
$$\langle \texttt{AmericanOpera} \sqsubseteq \texttt{Opera} \sqcap (\exists\texttt{ConductedBy.European}), 0.8\rangle$$

Suppose we are interested in those images that are about an opera conducted by a European director. To this end, we can use the query $\exists\texttt{About.}(\texttt{Opera} \sqcap \exists\texttt{ConductedBy.European})$. It can be verified that the degree of relevance attributed to `i` is 0.8, whereas that of `j` is 0.7.

References

[1] Franz Baader and Philipp Hanschke. A schema for integrating concrete domains into concept languages. In *Proceedings of IJCAI-91, Intern. Joint Conference on Artificial Intelligence*, pages 452–457, Sydney, 1991.

[2] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu. The Virage image search engine: An open framework for image management. In *Storage and Retrieval for Still Image and Video Databases IV*, volume 2670 of *SPIE Proceedings*, pages 76–87, San Jose, CA, February 1996.

[3] Alexander Borgida. Description logics in data management. *IEEE Transactions on Knowledge and Data Engineering*, 7:671–682, 1995.

[4] Martin Buchheit, Manfred A. Jeusfeld, Werner Nutt, and Martin Staudt. Subsumption between queries to object-oriented databases. *Information Systems*, 1994.

[5] Giuseppe De Giacomo and Maurizio Lenzerini. Making $\mathcal{CATS}$ out of kittens: description logics with aggregates. In *Proceedings of DL-95, 4th International Workshop on Description Logics*, pages 143–147, Roma, Italy, 1995.

[6] F. M. Donini, M. Lenzerini, D. Nardi, and W. Nutt. The complexity of concept languages. In *Proceedings of KR-91, 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 151–162, Cambridge, MA, 1991.

[7] Didier Dubois and Henri Prade. *Fuzzy Sets and Systems*. Academic Press, New York, NJ, 1980.

[8] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, and W. Niblack. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3:231–262, 1994.

[9] V. N. Gudivada and V. V. Raghavan. Design and evaluation of algorithms for image retrieval by spatial similarity. *ACM Trans. on Information Systems*, 13(2):115–144, April 1995.

[10] V. N. Gudivada and V. V. Raghavan, editors. *IEEE Computer. Special Issue on Content-Based Image Retrieval*. IEEE, Sept. 1995.

[11] E. J. Guglielmo and N. C. Rowe. Natural-language retrieval of images based on descriptive captions. *ACM Transaction on Information Systems*, 14(3):237–267, July 1996.

[12] H. V. Jagadish, A. O. Mendelson, and Tova Milo. Similarity-based queries. In *Proceedings of the 14th Symposium on Principles of Database Systems*, pages 36–45, San Jose, May 1995.

[13] C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In *Proceedings of SIGIR-93, 16th ACM Conference on Research and Development in Information Retrieval*, pages 298–307, Pittsburgh, PA, July 1993.

[14] Carlo Meghini and U. Straccia. A relevance terminological logic for information retrieval. In *Proceedings of SIGIR-96, the 19th ACM Conference on Research and Development in Information Retrieval*, pages 197–205, Zurich, August 1996.

[15] A. Rosenfeld and A. C. Kak. *Digital picture processing*. Academic Press, New York, 2nd edition, 1982.

[16] Manfred Schmidt-Schauß and Gert Smolka. Attributive concept descriptions with complements. *Artificial Intelligence*, 48:1–26, 1991.

[17] A. F. Smeaton and I. Quigley. Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of SIGIR96, the 19th International Conference on Research and Development in Information Retrieval*, pages 174–180, Zurich, CH, August 1996.