

Workshop Programme

14:30 – 14:45 Introduction (Thierry Declerck)

14:45 – 16:00 First Oral Session (20 mn per paper + 5 mn for questions)

14:40 – 15:10 Simone Santini - Context-based Retrieval as an Alternative to Document Annotation

15:10 – 15:35 Gregory Grefenstette - Comparing the Language Used in Flickr, General Web Pages, Yahoo Images and Wikipedia

15:35 – 16:00 Judith Klavans - Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access

16:00 – 16:30 Tea/Coffee Break

16:30 – 18:40 Second Oral Session (20 mn per paper + 5 mn for questions)

16:30 – 16:55 Taha Osman - Semantic-based Expansion of Image Retrieval Queries

16:55 – 17:20 Dina Demner-Fushman - Combining Medical Domain Ontological Knowledge and Low-level Image Features for Multimedia Indexing

17:20 – 17:45 Bertrand Delezoide - Object/Background Scene Joint Classification in Photographs Using Linguistic Statistics from the Web

17:45 – 18:10 Wei-Hao Lin - Identifying News Broadcaster's Ideological Perspectives Using a Large-Scale Video Ontology

18:10 – 18:40 Paul Buitelaar - Text Mining Support for Semantic Indexing and Analysis of A/V Streams

18:40 – 19:00 Closing Discussion (Thierry Declerck)

Workshop Organizers

Thierry Declerck – DFKI, Germany

Adrian Popescu – CEA LIST & Télécom Bretagne, France

Allan Hanbury – Vienna Univesity of Technology, Austria

Judith Klavans – University of Maryland, USA

Table of Contents

Context-based Retrieval as an Alternative to Document Annotation <i>Simone Santini and Alexandra Dumitrescu</i>	1
Comparing the Language Used in Flickr, General Web Pages, Yahoo Images and Wikipedia <i>Gregory Grefenstette</i>	6
Semantic-based Expansion of Image Retrieval Queries <i>Taha Osman, Dhavalkumar Thakker, Gerald Schaefer, Thomas Geslin, Philippe Kiffer</i>	12
Combining Medical Domain Ontological Knowledge and Low-level Image Features for Multimedia Indexing <i>Dina Demner-Fushman, Sameer K. Antani, Matthew Simpson, George E. Thoma</i>	18
Object/Background Scene Joint Classification in Photographs Using Linguistic Statistics from the Web <i>Bertrand Delezoide, Guillaume Pitel, Hervé Le Borgne, Gregory Grefenstette, Pierre-Alain Moëllic, Christophe Millet</i>	24
Identifying News Broadcaster’s Ideological Perspectives Using a Large-Scale Video Ontology <i>Wei-Hao Lin and Alexander Hauptmann</i>	31
Text Mining Support for Semantic Indexing and Analysis of A/V Streams <i>Jan Nemrava, Paul Buitelaar, Vojtěch Svátek, Thierry Declerck</i>	37
Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access <i>Judith Klavans, Carolyn Sheffield, Eileen Abels, Joan Beaudoin, Laura Jenemann, Jimmy Lin, Tom Lippincott, Rebecca Passonneau, Tandeep Sidhu, Dagobert Soergel, Tae Yano</i>	42

Author Index

Abels E.	42
Antani S. K.	18
Beaudoin J.	42
Buitelaar P.	37
Declerck T.	37
Delezoide B.	24
Demner-Fushman D.	18
Dumitrescu A.	1
Geslin T.	12
Grefenstette G.	6; 24
Hauptmann A.	31
Jenemann L.	42
Kiffer P.	12
Klavans J.	42
Le Borgne H.	24
Lin J.	42
Lin W.-H.	31
Lippincot T.	42
Millet C.	24
Moëllic P.-A.	24
Nemrava J.	37
Osman T.	12
Passonneau R.	42
Pitel G.	24
Santini S.	1
Schaefer G.	12
Sheffield C.	42
Sidhu T.	42
Simpson M.	18
Soergel D.	42
Svátek V.	37
Thoma G.E.	18
Thakker D.	12
Yano T.	42

Context-based Retrieval as an Alternative to Document Annotation

Simone Santini, Alexandra Dumitrescu

Escuela Politécnica Superior
Universidad Autónoma de Madrid

Abstract

This paper is a theoretical analysis of formal annotation and ontology for the expression of the semantics of document. They are found wanting in this respect, not only for technical reasons, but because they embody a fundamentally misunderstood model of the process of signification. We propose an alternative model in which the interpretation context plays a fundamental role in the definition of an *activity game* that includes all actions performed on a document, including accessing external data. We briefly discuss it and its current technical embodiment.

1. Introduction

After roughly ten years of work on image data bases, image semantics is, quite truly, the *bestia nera* of the field, the embarrassing skeleton in the closet of any researcher, the little dark family secret we don't like to talk about. We all know—unless we are kidding ourselves—that very little of what we do makes sense unless it can help somebody make semantic sense of the images in a repository, that none of our features or indexing schemes will be very useful unless they can somehow help the process of signification. This much, we know. On how this can happen our ideas are, by and large, rather more confused.

We have argued for some time that, in our opinion, the difficulties in extracting meaning from the images are due to the rather misled viewpoint that meaning is there in the first place. That is, our difficulties are due, more than to technical obstacles (which are also present and rather formidable but against which we are, all in all, well equipped) to a misunderstanding of the nature of meaning. One evident sign of this misunderstanding is the expression “extracting” meaning (an expression that we just used, in what we admit to have been a provocation), which subscribes to the essentialist view according to which meaning is something that exists *a priori*, behind the image, so to speak, and of which the image is but a code. If we could solve the problem of interpreting this code, the argument goes, technically difficult as this problem might be, we could “read” the image in the correct way, and have access to the *essence* of the meaning behind it.

This is, as we have come to realize, not the case, and the essentialist view is, if not unbearably naïve, at least desperately inadequate. Images are a node in a complex network of signification that goes beyond their content and includes other forms of textuality that go around them, as well as the cultural practices of the community that creates or receives images. There is, on other words, no meaning in images (or in anything else, for that matter) independent of the process of interpretation, a process that always takes place in a given context and as part of a given human activity.

These considerations extend to the relation between images and words. There isn't one relation between text and images but, rather, a multitude of modalities: text can

be used to explain images, images to explain text, images can set the mood in which text should be read, text and images can be two independent examples of the same category, text can contrast or contradict images, and so on. In this case as well, the most important thing that we should consider to make sense of the juxtaposition of text and images are the context of the search and the discursive practices of the environment in which the *sinoptic text* (the organized layout of text and images) was created.

2. Visualism Nailed

One of the questions that the call for paper presented to the participants to this workshop was: “what elements in a lexicon correspond to picturable objects?”. This question reveals, in a rather transparent way, a presupposition that underlies the whole area of image annotation¹: images contain *objects* and, somehow, the meaning of an image is a function of the objects it contains. Sometimes this hypothesis is strengthened to encompass compositionality: the meaning of an image is a function of the meaning of the objects it contains. With a certain flair for simplification, one could say that this point of view endorses statements such as “the image of a pencil means ‘pencil’”, or “the image of a nail means ‘nail’”. If somebody truly believes this, we see no better way to dispel it than reporting an example from a professional: the Mexican photographer Pedro Mayer². In response to the question “would not a photograph of a pencil on a table always be just that, a photograph of a pencil on a table?” he replied:

[Researchers in Peru had] the idea of using cameras to discover the codes being used by [poor Peruvian children]. They would ask a simple question and then elicit from the children [...] a response with a picture made with very simple cameras.

They wanted to know what these children thought of “exploitation” [...]

One child came back with a picture of a nail on a naked wall. At first the instructors thought that the child had misunderstood the

¹ Some people prefer to use the charming but etimologically incorrect term meta-data.

² The example was reported in (0), and we used it before in (0)

idea of the experiment. But upon further investigation they found out that these children were living in an extremely poor town, several miles outside of Lima, and in order to make a little bit of money they walked every day all those many miles into town to shine shoes. And so that they did not have to carry back and forth the heavy load of the shoe boxes they rented a nail on the wall at someone's place in town. And the man that rented the nail charged them for this half of what they earned for the whole day. As you can see, sometimes a nail on the wall means much more than a nail on the wall. Or for a Cuban child the picture of a pencil on a table might have implications dealing with the blockade, as they had no pencils for a long time.

This example is interesting because, all in all, the contents of the image are quite irrelevant to their meaning. Better yet: the contents of the image are relevant only inasmuch as they are apparently unrelated to their meaning, and assume a relevance once they are transformed by their intended meaning. This is an important reversal of the naïve assumptions about meaning: it is not the contents of the image that determines its meaning; it is the meaning (given, in this case, by the material circumstances of production and by the discursive practices that guide the process of production of photographs) that grabs the contents, repossesses them, and uses for its own semantic purposes. If we want to discuss this image, we should not ask ourselves what it contains but, rather, what instruments of communication have been used for its production, what relation between objects and states of the world have been singled out by the choice of this particular content, and what discursive practices make this relation an acceptable way of communicating. In this case, a relation of metonymy leads from an object (a nail on somebody's wall) to the exploitative relation of which it is a part. The relation between the image and its meaning is given here by a particular context: that of its production. We can't really understand the image unless we are told the story of its production and of the context in which this happened.

These considerations (of a much wider generality than this simple example) deny validity to what might be called the naïve theory of annotation, according to which describing the objects in an image and their relations reveal its meaning. One can still defend annotation, though: it is still possible, the argument would go, to fathom a system that would formalize not only the contents of an image, but also its relations with the textual, iconic, and iconographic elements that surround it, the discursive practices of its creation and (but here we are stretching plausibility) the complex relation between contents, context, and discursive practices on one side and meaning on the other. We believe that, even in this somewhat more sophisticated setting, the trust in annotation is misplaced for two orders of reasons: the dependence of meaning on the interpretation process, and the existence of that pesky inconvenient called human nature, which, as much as computing scientists are keen

on ignoring it, keeps popping up whenever people are involved in the use of computers. We will devote the following two sections to these phenomena, starting with the latter.

3. Human Nature

To the best of our knowledge, the only call for common sense in the orgiastic euphoria about the possibility of consistent and correct annotation of content came from Cory Doctorow (0), whose arguments—some of them—we will briefly sketch in this section. Doctorow considers seven reasons why “meta-data”, even if they were conceptually possible, would not work in actuality; some of these reasons are not particularly relevant in this context, so we will concentrate on the relevant ones.

First pragmatic matter: *people lie*. People who want you to look at their content (companies, for instance) will write all sorts of falsehood to attract you. By and large, writing one or two falsehood is called lying, writing many big ones is either marketing or politics. You will find many examples of both in any annotation corpus not produced under strict control. That people lie to make you look at their page is the reason why every day we receive one or two email saying that a nice, bored girl is dying to meet us, or that somebody would like to use our bank account in order to transfer 30 million dollars out of Uganda.

Second problem: *people are lazy*. If people are not interested in using annotations/titles/notes to lie, they are probably not interested in using them at all. So, your electronic mail is full of messages without a subject line, or with a subject line such as “Re: X”, where X is the subject line of a message that was sent three months ago; disks are full of documents called “untitled.doc” and so on. In other words: people can be counted upon only to create the annotations that are useful to them, but they will not spend any amount of time, no matter how small, to create annotations that are useful to you.

Third matter: *people are stupid*. As a matter of fact, people can't even be trusted to be thorough when their own interest is at play. Seven years after Doctorow's example, you can still find a sizeable number of PDAs and bonzai trees on e-bay by typing the word “plam”. If people can't take the time to correct their spelling when their sales depend on it, how can you trust them when they are just producing annotations for your searching convenience?

In addition to these human, so to speak, problems, Doctorow highlights some issues related to the nature of formal, hierarchical annotations. In particular, hierarchies are never neutral; they derive from a certain value system, and the semantic axes along which the divisions are made, or the order in which they are made, are an expression of this system. It is not possible to use a classification without accepting its value system. A car taxonomy may start, at its highest division level, with the type of the car (sports, sedan, hatchback,...) or with a mileage classification (low, medium,...). We suspect that greenpeace would propose to use the second, and that General Motors would opt for the first. The point is: neither or them is innocent or neutral. Each classification represents an ideological commitment that is forced upon the user.

4. The Death of the Reader

Even if we assumed that formal annotation could do incomparably more than listing the objects in an image, even if we assumed that people were willing to be thorough and truthful, the final considerations of the previous section evidence a further—fatal, in our view—limitation of annotation, be it formal or not: annotation always embodies a normative notion of meaning, one that takes in no account the process of interpretation and the circumstances in which it takes place. Roland Barthes (0) criticised the habit of interpreting a text in reference to the (hypothetical) intentions of its author: what we are really faced with is an autonomous text and an interpretative situation. Whether there was an author, what the author thought, what he believed, what he intended to communicate, is simply a speculation that we make during the interpretation act: the personality of the author is created jointly by the text and by the community of reader, his intention is a model that the reader uses in order to create a sense for the text. It is an act of interpretation, and a text would not change its meaning if we discovered that the author didn't exist at all, as it happened for the Ilyad and the Odyssey. Barthes calls this discovery the “death of the author”.

The field of formal annotation seems to have taken the opposite point of view: far from proclaiming the demise of the author and the opening of the possibilities of interpretation, it tries to normatively lock the “intended” meaning in a formal structure, to make signification independent of the act of reading and of the circumstances in which reading takes place. Annotation attempts to freeze the free play of signification, fixing it once and for all and making it independent of the reader. In its most explicit incarnations, such as the semantic web, the formalization of meaning is pushed to a point where the reader can be disposed of completely: the formal specification of meaning can be read by an algorithm. To the Barthesian death of the author, annotation responds with a rotund call for the death of the reader. But if the reader is dead, there is nobody left to make sense of a text, and reading is always reading in a given context. To translate these observations into a more explicit form, something closer to the computational needs, one might say: there can be no semantic (whatever that might mean) access to information unless we take into account the context in which access is made.

5. Context based Retrieval

The previous considerations seem to point towards a certain number of criteria for the design of an image retrieval system:

- I. To have a model of content, however determined, is not sufficient for retrieval; the most important model is that of the context in which search takes place. Modelling content is not a value per se, but it is only valid as a way of interpreting content from the point of view of context.
- II. Images do not signify alone; they acquire meaning, as everything else, from the context in which they are accessed. Additionally, they acquire meaning from the relation with the elements with which they appear, from the

meaning of these elements, and from the discursive practices that constraints the acceptable ways in which they can be put together.

- III. In any case, one should not trust annotation, be them in the form of labels, of natural text specifically written for the purposes of facilitating search, or in more formal ways such as predicates and logic theories (also called “ontologies”). It might be too much to say that all people are lazy and stupid liars, but it is certain that whatever they write in the way of annotation (if anything) can't quite be trusted. The only reliable source of information about the data is the data themselves.

The first point is technically the most relevant of the three because the techniques taht we develop to formalize context and to determine the relation between two contexts can then be applied to the characterization of the production context in which an image is placed, and because succeeding in such a characterization without the intervention of special annotation would lead to a solution for the third problem.

Before continuing, reliance on context must somehow be justified, and the relation between context and meaning must be, in part at least, clarified and formalized. The relation that is generally assumed between meaning and text (we are using the word “text” in a very general sense here: any thing that signifies using an established code is a text in our definition) is that the context contributes to the interpretation of the text that is, in a somewhat simplistic way, that context is that thing which changes the meaning of a text.

A useful point of view is to reverse this definition: texts, once assimilated or spoken, change the context in which they are placed, and the meaning of a text is precisely the change of context that is provokes. Meaning is to be sought in the relevant context changes that can take place in a given communicative action. Consider the following example (0): I am asking “does Maria lke wine?” and you answer me “She hates all kinds of alcohol”. What is the meaning of your statement? I am into a context in which I don't know whether Maria likes wine or not, and I am seeking a movement into a context in which either Maria likes wine or she doesn't. The meaning that we give to the answer is the one that provokes one of the desired changes in context. In this case, I will interpret your general statement about Maria's dislike for alcohol as saying that she doesn't like wine. Formally, if one has a context C and the presence of a text t changes it to a context C' , the meaning of text t in that situation is $\mu(t) = \Delta(C, C')$.

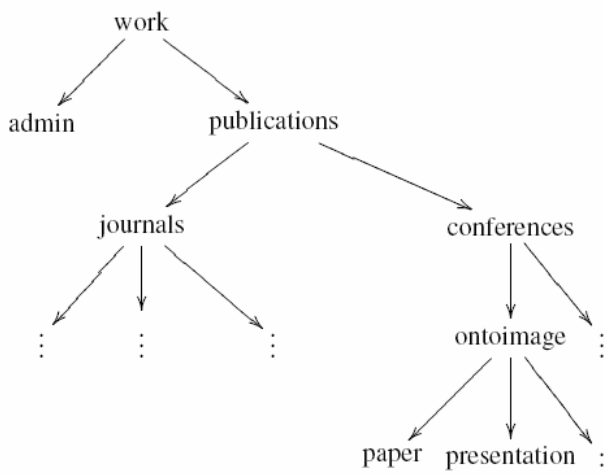
This point of view, when applied to data semantics, changes completely our way of looking at the problem. For one thing, one can no longer talk about teh meaning of a document: meaning depends on the context to which the document is applied. If we want to continue to talk about the meaning of a document, te only way in which we can (approximately) do so is by regarding it as a function that transforms contexts into contexts that is, in mathematical terms, the meaning of a document is an endorphism in the context space:

$$\mu(t) : C \rightarrow C(1)$$

This point of view also helps us answer an important foundational question: can meaning be formalized? It should be obvious that the question, thus as it is posed, does not have an answer: meaning can only be formalized to the extent that the context to which it applied can, and only when applied to formalized contexts. In a completely free and unconstrained context, there is very little we can do to formalize meaning. In this sense, the dream of the semantic web to formalize meaning so that anybody, in any context, can do “semantic data access” is an illusion. In more structured activities, in which context can, in part, be formalized, the formalization of meaning comes naturally. This point of view even leaves some space for annotation. If you still believe that people are reliable annotators, then you can fathom a way to describe the discursive practices that controlled the production of a document and, possibly, the how and why of the relations between different parts of it. This would not, by any means, be the meaning of the document, but it would formalize the *production context* that could then interact with the *reception context* of the reader (the one that we called simply “context” above) to change it: the change in the reception context caused by the text-cum-production-context is meaning. We will not pursue the point further in this paper.

6. Getting the Context

In order to carry out the programme outlined in the previous section it is necessary to know the context of a data access, and to have it available in a machine processable form. There are many human activities in which this is not possible: when you play poker with a group of friends or go out with your girlfriend (resp. boyfriend) there are no processable traces of the meaning of these activities: they leave no digital trace. On the other hand, for many of us, many work activities are performed using a computer, and they do leave a digital trace that can tell us about their context. Suppose we are preparing a presentation for a conference to which we had submitted a paper and that, during this process, we need to clarify a point or to look for an illustration for our presentation. In order to prepare our presentation, we have created a document in a directory (let us say the directory presentation) where we have possibly copied some documents that we thought might be useful. This directory is likely to be placed in a hierarchy, something like this:



Its sibling directories will contain documents somehow related to the topic at hand although, probably, not so directly as those that can be found in the work directory. The siblings of the conference directory (and their descendants) will contain documents related to our general area of activity, although not necessarily directly related to the topic of the presentation. This project, in its context search component, will look for ways to use this information in order to direct and focus the search. This information will constitute the context of the search. One consequence of this point of view is important enough to be noted from the outset: data access is no longer an independent activity, but can take place only in the context of a certain activity.

In section 2 and in point 3 of section 3, we chastized, quite severely, the reliance on users to annotate what they are doing, based on the general lazyness of the average computer user (we could paraphrase Barnum and say that “nobody has ever lost money for overestimating the lazyness of the computer user”). The same general objection would appear to apply to the use of the directory structure as an indicator of relevance. We all know somebody whose desktop is completely covered with documents because that person has never created a directory in his life. In part this objection is valid, but there are two considerations that suggest that, in the case of context, the problem might not be as serious as in the case of annotations:

- I. In the case of annotations, people have no incentive to annotate other than making somebody else’s life easier (namely, making life easier for the person who searches); in the case of directory creation, the creator is making a service to himself, since directories are useful, independently of searches, in order to organize one’s work. Because of this, we can expect that lazyness will be less of a factor for the creation of directories than it is for the creation of annotations.
- II. In any case, a rich directory structure will help us formalize the context, but formalization does not depend entirely on it. That is, even in the absence of a good structuring of directories, we can still gather information about the context of a data access.

Utilitarianistically, one can say that the structuring and the discipline that leads to a good context formalization is useful for the person who makes the searches, while the discipline that leads to reliable annotations is useful mainly to people other than the annotator. Human nature being what it is, we can expect that the first form of effort will gain more adepts than the second.

6.1 Context Representation

The problem of representing context and, most importantly, to make it interact with the documents, is still largely unexplored, at least as far as computing science goes, and it is not clear in which direction one should look for a proper representation.

As a first step, one might consider the use of techniques

from information retrieval. Here, we will give some pointers on the possible use of a model based on a vector space representation of word contexts (0), and a self-organizing map to give a non-linear form of latent semantics. This is the model that we are currently using in our activity. Word context are groups of sequential words that occur in a text. In this case, they are more representative than single words because they capture, statistically, collocations, which are relatively strong indicators of the meaning that a word is given a certain context (if the word “bank” cooccurs with “investment”, it is likely to mean a financial institution, if it co-occurs with the word “river”, it is likely to indicate the border of a body of running water, and so on). These co-occurrences will be represented in a suitable feature space and a *self-organizing map* (0) will be used to cluster and represent its contents, using again fairly standard techniques (0).

The map represents a sort of non-linear latent semantic subspace: it captures the statistically significant relations between terms in a given context. The learning algorithm gives us an obvious way to start including the structure of the directories into the context representation: learning may not be limited to the documents contained in the working directory, but can include those contained in the children/siblings/parents; furthermore, by varying the fraction of times the documents in each one of these directories are presented we can give more or less importance to certain structural relations. Finally the map, being geometric in nature, suggests a way to extend the context representation to multi-media document or, at least, to documents containing images. We can extract image features that can be represented in geometric spaces (0) and derive the direct sum of the space of words and the space of features. This should allow the map to capture any statistical regularity, in the document corpus, between certain word combinations and certain image features. Note that if the features are extracted from regions of the image, one can use feature context techniques similar to the word context used for text, thus seeking statistical regularities between co-occurrences of words, and co-occurrences of localized features.

While the techniques used in this approach are fairly standard, its novelty is that, in this case, we are not using them in order to represent the data base in which the search is to be done, but to represent the environment from which the query originates.

7. Words of Parting

We have argued that formal annotation, and the general ontological programme that comes with it, might not be the proper way to consider the problem of the meaning of the data and, in general, to frame the issues related to semantics.

This position goes against a certain common sense philosophy. We can look at texts, read them, and make sense of them, and it seems natural to interpret this act as unlocking the meaning that is in the text. After all, if we don't know from what gate does flight 354 to New York leave, and we read the announcement board of the airport, we end up knowing it. It is easy to model this situation as a transfer of a specific information (viz. that the flight leaves from gate C34) from the announcement board to me. The error is the failure to recognize that this is a limit

case, namely a case in which the external context is so constraining that the reading of the symbol “C34” can basically have only an interpretation, and to extend the same model to the common situation, the one in which the interpretation context plays a much more important rôle. We have given our arguments (convincing, we hope) why we believe that this position represents a gross philosophical simplification, and we believe that it will ultimately result in the sterility of semantic computing. Technically, this paper has presented the outline of a different model of meaning, one in which the reader's context plays a preponderant rôle. We have presented a simple framework that in the future will be extended in different directions: on the one hand, the integration in this framework of more formal representations, at least for those parts of the context that can be formalized; on the other hand, the development of suitable data base techniques to make this kind of query efficient.

8. References

- Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 252--262.
- Roland Barthes. *S/Z*. Paris:Seuil, 1976.
- Cory Doctorow. *Metacrap: Putting the torch to seven straw men of the meta-utopia*. Web published, 2001.
- S. Kaski. Computationally efficient approximation of a probabilistic model for document representation in the WEBSOM full-text analysis method. *Neural Processing letters*, 5(2), 1997.
- T. Kohonen. *Self-organizing maps*. Heidelberg, Berlin, New York: Springer-Verlag, 2001.
- S. Laakso, J. Laaksonen, M. Koskela, and E. Oja. Selforganizing maps of web link information. *Advances in Self-Organising Maps*, page 146-151, 2001.
- Fred Ritchin. *In Our Own Image*. Aperture, 1999.
- José Carlo Rodríguez. *Jugadas, partidas y juegos de lenguaje: el significado como modificación del contexto*. Asuncion: Centro de documentos y estudios, 2003.
- S. Santini. *Exploratory image databases: context-based retrieval*. San Diego: Academic Press, 2001.
- Simone Santini. Image semantics without annotation. In S. Handschuh and S. Staab, editors, *Annotation for the Semantic Web*. IOS Press, 2003.

Comparing the Language Used in Flickr, general Web Pages, Yahoo Images and Wikipedia

Gregory Grefenstette

CEA LIST

18, rte du Panorama, BP6, Fontenay Aux Roses -F-92265 France

E-mail: gregory.grefenstette@cea.f

Abstract

Words can be associated with images in different ways. Google and Yahoo use text found around a photo on a web page, Flickr image uploaders add their own tags. How do the annotations differ when they are extracted from text and when they are manually created? How does these language populations compare to written text? Here we continue our exploration of the differences in these languages.

1. Introduction

Language models associate frequency to language phenomena (Croft and Lafferty, 2003). Unless images are being matched up using low-level visual features (Vasconcelos, 2007), words are used to index and retrieve images. It is important to know what words are used to index images and what the characteristics of these words are. The principal characteristic that information retrieval uses is word frequency, a simple uniterm language model. In this article, we present a preliminary exploration of the types of words used to index images on the web, using the uniterm language model of image index terms in two different image collections: Flickr images and Yahoo Images. These language models are compared to the models found in two text-based web collections: Wikipedia, a clean, edited collection, and in general web pages indexed by a general search engine, Exalead.

Yahoo Images extracts index words from metadata found around images: filenames and text in links towards pictures, as well as from surrounding text (Datta *et al*, 2008). Flickr indexes pictures using user-supplied words (Marlow *et al*, 2006). We will show below that these two language populations are different, with Yahoo language often resembling edited text such as that found in Wikipedia.

2. Related Work

Hanbury (2006) did a study of the words that were used in image annotation experiments. These sets were all restricted to a few hundred words maximum, since the purpose of these experiments was to assign low-level visual features to one or more of these concepts.

Little work has been done on studying the specificities of human supplied tags for image tagging (Ji *et al*, 2007). What is this uncontrolled language, how does it differ from uncontrolled written text? We have begun examining the differences between the general language models of web-based text and the unrestricted tags human add to images. We found (Pitel *et al*, 2007) that there is a

correlation between co-occurrence statistics of animal names and background scenes in running text, and in human supplied tags. We also found differences (Grefenstette and Pitel, 2008) between the types of words used to index images by hand and words extracted from text, for example there are more nouns used in hand image tagging than adjectives, compared to the proportion used in written text. Certain activities are overrepresented in image tags (e.g. weddings) than in written text. This article continues these preliminary investigations, examining additional language sources: Wikipedia and Yahoo! Image indexes.

3. Different Language Models

In order to compare different language models, we began with a list of English words, drawn from a full form lexicon of English words. Over a period of time, each word was submitted to a search engine and the web pages frequencies of these words were recorded. In order to anchor the web search in only English pages, to avoid overcounting cognates from different languages (for example, conflating counts of the English and German word *die*) we recorded the page frequencies of pages containing the word plus a number of common English words, so the query for the word *die* would look like “die” “the” “with” “and” “in” “of” (Grefenstette, 2007). We sorted the words by their frequencies, and retained 5000 frequent words more than three letters long. This led to a list of words such as *abandoned*, *ability*, *able*, *above*, *abroad*, *absence*, *absolute*, *absolutely*, *abstract*, *abuse*, *academic*, *academy*, *accept*, *acceptable*, *acceptance*, *accepted*, *access*, *accessibility*, *accessible*, *accessories* ...

We used this list as a language sample to compare usages in four different collections: (i) pages indexed by Exalead, a French search engines that has indexed 8 billion URLs, (ii) Flickr tags, (iii) Wikipedia web pages indexed by Google, and (iv) Yahoo! image search. The Exalead counts which index all text present on the web pages, gives an idea of the frequency of word use in written text. The Wikipedia counts give an idea of language use in a more controlled, edited environment

than the entire web. Flickr tags counts show which words human users choose to annotate their photos. Yahoo counts gives us the counts of both words appearing in text near images, as well as words appearing in image file names (e.g., *dog.jpg*) that are the results of a human decision.

3.1 Exploring the differences

In order to explore the differences in the language use in the four sets, we begin by looking at the most common words in each set. The following lists are ranked by frequency from the most frequent word (out of our subset of 5000 web-frequent English words) to less frequent. The ten most frequent words in each index are shown below:

<i>Ten most frequent words from our 5000 words</i>			
Exalead	FlickrR	Wikipedia	YahooImages
people	wedding	random	image
good	party	views	index
work	travel	foundation	full
need	family	modified	photo
used	Japan	powered	images
know	vacation	interaction	photos
read	London	encyclopedia	size
great	beach	discussion	gallery
available	friends	user	modified
want	trip	create	small

From this cursory examination, we see that Exalead (web text) contains many general words; that Wikipedia (encyclopedia articles) and Yahoo Images (images indexed using words near images or in file names) contain mostly metadata words corresponding to the type and structure of data that they index; and that Flickr (hand given photo tags) describes images, showing the most popular images with tags that one might expect.

If we skip the first 100 most common words in each set, in order to skip the metalanguage in the Wikipedia and Yahoo Image sites, we find these words starting from the 101st most common word, and we see that Flickr and Yahoo both have place names (proper and common nouns) and objects, while Wikipedia and Exalead contain more technical terms:

<i>Words in ranks 100 to 110th most frequent</i>			
Exalead	FlickrR	Wikipedia	YahooImages
access	photos	military	return
technology	tour	people	little
questions	football	kingdom	good
tools	Asia	shown	media
phone	band	place	island
times	yellow	need	history
making	June	thanks	right
national	August	police	June
subject	bird	Frank	read
office	dance	November	holiday

Our list of 5000 frequent English words contains 364 geographical and personal single-word names, taking this list, we see that the web, Wikipedia and even Yahoo

Images to a large extent, are concerned with dates and places, while Flickr users choose country names very often to tag their pictures.

<i>Ten most frequent uppercase initial words</i>			
Exalead	FlickrR	Wikipedia	YahooImages
International	Japan	Charity	Sale
English	London	March	International
Sale	Italy	America	York
March	France	Europe	April
America	Paris	English	English
June	China	Africa	Japan
April	Europe	Asia	John
August	Australia	February	Street
December	Canada	December	July
York	Germany	Frank	June

3.2 Semantically typed differences

We found it interesting in (Grefenstette and Pitel, 2008) to compare semantically distinct word classes from different language populations. If we just take country names, for example all the one-word country names found in Wikipedia entry “List_of_countries”, we find that the ranking of countries (the US has been excluded from the list since it varies in its spelling) resemble each other, but when we look at the average rank of country names (the last row), we see that country names are much more prevalent in Flickr tags (where the average rank of the ten most popular country names is 25) than in the other sources. They are among the favourite tags of Flickr users who supply their own tags.

Exalead	FlickrR	Wikipedia	YahooImages
Canada	Japan	Germany	Japan
Australia	Italy	France	France
Spain	France	Australia	Canada
India	China	Japan	China
Russia	Australia	Canada	Australia
China	Canada	Senegal	Italy
Turkey	Germany	Indonesia	India
Israel	Spain	Italy	Mexico
Italy	India	India	Germany
Japan	Taiwan	Ireland	Spain
<i>Average rank for 10 most popular countries (N=5000)</i>			
1382	25	787	205

Beyond lists of country names, we can find other lists of things on the Web. For example, in the lexical hierarchy WordNet³, we find the synset *sport*, *athletics* (an active diversion requiring physical exertion and competition), which is considered a hypernym for 116 terms including: *acrobatics*, *angling*, *aquatics*, *archery*, *athletics*, *badminton*, *ball*, *ballgame*, *baseball*, *basketball*, *bathe*, *battledore*, *battue*, *beagling*, *bicycling*, *bobsledding*, *boxing*, *bullfighting*, *cast*, *casting*, *cockfighting*, *coursing*, *crab*, *cricket*, ... The most common of these sport terms are given below. Again, by looking at the average rank of the ten most common sports terms, we see that Flickr

³ <http://wordnet.princeton.edu/>

users often choose sports names as tag. The words found around images by Yahoo, also contain more sports names than in text found in Wikipedia or on the web by Exalead.

Exalead	FlickrR	Wikipedia	YahooImages
racing	football	soccer	golf
basketball	baseball	running	racing
swimming	soccer	football	football
singles	racing	singles	fishing
football	hockey	baseball	soccer
riding	skiing	basketball	baseball
hunt	cycling	hockey	running
hockey	basketball	racing	basketball
surfing	fishing	athletics	tennis
diving	diving	hunt	hockey
<i>Average rank for 10 most popular sports (N=5000)</i>			
1827	229	1599	612

Leaving aside country names, associated with tourism, and thus pictures, and sports events, we can examine other semantic domains. If we take the Roget's entry 737b⁴ for POLITICS, we find 75 one-word phrases in our list of 5000 frequent words, including: *Abolitionist, activist, also-ran, aspirant, backer, bailiwick, ballot, campaign, campaigning, candidacy, candidate, communism, Communist, conservatism, Conservative, constituents, contributor, democracy, Democratic, election, electioneering, electorate,...* The last line of the table below shows that these political terms are more highly ranked in Wikipedia than in Flickr or Yahoo images.

Exalead	FlickrR	Wikipedia	YahooImages
program	party	opinion	party
party	politics	official	official
voice	Labor	issue	program
issue	labor	position	stand
official	election	party	issue
political	campaign	program	politics
opinion	program	political	voice
position	stand	politics	position
campaign	democracy	vote	winner
politics	platform	stand	campaign
<i>Average rank for first 10 political terms (N=5000)</i>			
523	1251	566	967

If we take single-word color terms⁵ we find many words that are both colors and objects (silver, brass, chocolate, rose) and we cannot distinguish the senses of the words by our simple counting techniques, but we see by the ranking of the color tags in Flickr and words associated with images in Yahoo, that color names appear more often than in general text. Below in the next section, we will see that this popularity does not hold for all adjectives, so we should consider that people use colors just because they are describing the characteristics of their images.

Exalead	FlickrR	Wikipedia	YahooImages
---------	---------	-----------	-------------

⁴ Roget's 1911 version can be found at, for example, http://www.abcd-classics.com/rogets/rogetsthesaurus/rogets_body-0737.html

⁵ http://en.wikipedia.org/wiki/List_of_colors

red	blue	red	black
white	red	black	red
blue	white	white	white
yellow	black	blue	blue
black	yellow	brown	silver
brown	pink	silver	rose
chocolate	rose	rose	yellow
silver	brown	yellow	pink
brass	sepia	grey	brown
salmon	rust	bronze	chocolate
<i>Avg. rank for 10 most popular colors (N=5000)</i>			
1490	200	1100	365

Another set of common words are names of animals⁶. We see that animal names appear more frequently in Flickr tags than either in Wikipedia (where all animals are nonetheless all described) or on the web in general. Once again the counts in YahooImages falls somewhere between Flickr and Wikipedia.

Exalead	FlickrR	Wikipedia	YahooImages
bird	bird	bear	fish
mouse	fish	fish	bird
chicken	bear	bird	bear
fish	monkey	mouse	tiger
wolf	spider	eagle	eagle
eagle	tiger	wolf	mouse
salmon	chicken	seal	wolf
crab	eagle	tiger	monkey
bear	mouse	monkey	spider
dove	crab	chicken	chicken
<i>Avg. rank for 10 most popular animals (N=5000)</i>			
2731	678	2586	1231

A semantic set which is more common in Wikipedia than in Flickr is career names⁷. Wikipedia contains many biographical articles, so it is not surprising that the frequency of these job titles is more common than on the general web (Exalead).

Exalead	FlickrR	Wikipedia	YahooImages
artist	model	editor	model
editor	photographer	director	artist
doctor	artist	writer	student
secretary	student	producer	designer
teacher	cooper	artist	photographer
attorney	cook	student	director
guard	designer	model	driver
judge	actor	actor	teacher
model	mason	minister	professor
director	baker	professor	cook
<i>Avg. rank for 10 most frequent careers (N=5000)</i>			
1345	967	871	953

Common place names⁸ appear often in Flickr tags, but most often in YahooImages, maybe because of online stores (with indexed images).

⁶ http://en.wikipedia.org/wiki/List_of_animal_names

⁷ <http://www.sff.net/people/julia.west/CALLIHOO/jobs.htm>

⁸ <http://www.sff.net/people/julia.west/CALLIHOO/places.htm>

Exalead	FlickrR	Wikipedia	YahooImages
house	beach	school	club
office	park	club	park
store	house	town	house
shop	church	university	hotel
school	museum	house	shop
market	school	building	school
club	building	library	beach
library	club	park	university
park	boat	hall	store
hall	castle	office	museum
<i>Avg. rank for 10 most popular places (N=5000)</i>			
265	93	517	55

Although images can easily convey emotion (Junghoefer *et al*, 2001) emotive state words⁹ appear more often in Wikipedia and the Web than on image tags. If people one considers that people do not describe the obvious, then why are color words often used as tags, and not emotions?

Exalead	FlickrR	Wikipedia	YahooImages
happy	happy	interested	sharing
secure	warm	happy	happy
understanding	calm	strong	warm
glad	strong	satisfied	strong
satisfied	angry	understanding	secure
comfortable	sharing	confused	interested
worried	scared	concerned	inspired
strong	hurt	glad	comfortable
scared	loving	inspired	loving
sharing	inspired	afraid	understanding
<i>Avg. rank for 10 most frequent emotive words (N=5000)</i>			
1705	2131	1333	2112

Surprisingly food¹⁰ is a very popular tag in Flickr. There are many pictures of food. Maybe as part of tourist excursions, people take pictures of restaurant meals.

Exalead	FlickrR	Wikipedia	YahooImages
coffee	fish	fish	fish
cookies	apple	rice	apple
chicken	cake	apple	coffee
apple	coffee	fruit	cake
fish	chocolate	coffee	chocolate
bread	fruit	cherry	fruit
candy	pumpkin	milk	candy
chocolate	candy	chicken	cherry
rolls	pizza	cheese	chicken
chips	cherry	chocolate	rice
<i>Avg. rank for 10 most popular food (N=5000)</i>			
2223	519	3353	1274

Pastimes and hobbies¹¹ are also very popular image tags:

Exalead	FlickrR	Wikipedia	YahooImages
---------	---------	-----------	-------------

⁹ <http://www.sff.net/people/julia.west/CALLIHOO/emotions.htm>

¹⁰ <http://www.mrsjonesroom.com/jones/foodalphabet.html>

¹¹ http://en.wikipedia.org/wiki/List_of_basic_hobby_topics

music	music	soccer	music
writing	camping	writing	photography
dance	football	music	golf
photography	dance	football	football
literature	photography	literature	dance
basketball	soccer	dance	fishing
drawing	dancing	basketball	painting
swimming	painting	aviation	camping
football	skiing	tennis	soccer
walking	cycling	drawing	walking
<i>Avg. rank for 10 most popular hobbies (N=5000)</i>			
1064	158	1372	369

3.2 Grammatically typed differences

In addition to semantic classes, we can also examine grammatical classes. Many abstract words in English end in -ation. There are 75 such words in our list of 5000 common English words. When we compare the ranks of these words in each set, we see that they are often used in Wikipedia. The first words in the Wikipedia list below correspond to Wikipedia page structure, but the effect of the relative disaffection for these abstract words in Flickr still persists down to the 50th more frequent word.

Exalead	FlickrR	Wikipedia	YahooImages
education	vacation	foundation	vacation
location	celebration	navigation	navigation
registration	station	documentation	education
application	aviation	location	location
association	education	population	association
foundation	foundation	station	station
communication	installation	association	accommodation
station	transportation	education	foundation
navigation	association	organization	installation
publication	inspiration	creation	presentation
<i>Average rank for first 10 terms in -ation (N=5000)</i>			
588	786	423	515
<i>Average rank for first 50 terms in -ation (N=5000)</i>			
1707	2263	1870	2101

Similarly with long words (10 letter words from our 5000 word collection), we find a persistent disaffection as image tags:

Exalead	FlickrR	Wikipedia	YahooImages
technology	university	foundation	university
conditions	restaurant	discussion	collection
newsletter	basketball	categories	networking
experience	motorcycle	registered	discussion
registered	conference	navigation	technology
management	engagement	copyrights	management
understand	exhibition	identified	navigation
individual	convention	references	restaurant
activities	tournament	considered	photograph
especially	industrial	experiment	conference
<i>Average rank for first 10 ten-letter words (N=5000)</i>			
273	411	54	280
<i>Average rank for first 250 ten-letter words (N=5000)</i>			
2388	3327	2523	3134

Although all four collections have many short words (here

we look at 5-character word) as indexes, we see here that those in Flickr image tags are descriptive of scenes.

Exalead	FlickrR	Wikipedia	YahooImages
great	party	views	image
years	beach	pages	index
right	music	check	photo
today	canon	tools	small
using	water	thank	album
place	night	users	music
order	green	avoid	forum
state	white	latin	party
music	house	album	video
check	trees	learn	house
<i>Average rank for first 10 five-letter words (N=5000)</i>			
20	33	48	13

The short words used in Yahoo image tags seem to come from metadata and online commerce.

From a large lexicon of English word forms (265,531 unique tokens), we extracted sublists of words which were unambiguously nouns or verbs or adjectives, and filtered this list through our set of 5000 frequent English words. The next three tables show the most common words found in each set for each language population.

Unambiguous nouns have a relatively even distribution in each group.

Exalead	FlickrR	Wikipedia	YahooImages
years	family	foundation	photo
products	friends	interaction	photos
music	nature	encyclopedia	album
user	music	discussion	description
events	canon	user	parent
family	sunset	categories	music
version	portrait	text	forum
health	food	events	family
history	architecture	version	hotel
software	lake	history	friends
<i>Average rank for 10 most frequent unambiguous nouns (N=5000)</i>			
49	27	17	18
<i>Average rank for the 100 most frequent unambiguous nouns (N=5000)</i>			
276	293	308	224

Unambiguous verbs are relatively rare (many English verbs are also nouns). In our 5000 frequent words only 35 were listed as ambiguous. These rare verbs are more common in the text sites, web (Exalead) and Wikipedia, than as image tags.

Exalead	FlickrR	Wikipedia	YahooImages
including	snorkeling	creating	including
includes	emerging	including	includes
requires	including	includes	creating
depending	creating	requires	knew
brings	tobogganing	selecting	brings

depends	happens	exists	happens
knew	campaigning	involving	bringing
happens	reflects	knew	requires
bringing	brings	depending	specializing
enables	knew	bringing	announces
<i>Average rank for 10 most frequent unambiguous verbs (N=5000)</i>			
1036	3581	1223	2925
<i>Average rank for the 30 most frequent unambiguous verbs (N=5000)</i>			
2581	4417	2557	4020

Adjectives are common in Web text, especially in edited text such as Wikipedia, and less popular in Flickr tags (color tags notwithstanding). This might seem surprising because adjectives are generally considered as descriptive, and one might expect that people adding tags to their photos would want to describe then using adjectives, but this does not seem to be the case. The adjectives that are used in Flickr are physically descriptive (compared to adjectives like *financial*, *virtual* found in other sets), but they are less often used than nouns.

Exalead	FlickrR	Wikipedia	YahooImages
available	cute	recent	previous
latest	sexy	available	larger
recent	happy	anonymous	untitled
important	outdoor	printable	available
previous	golden	reliable	happy
able	rural	able	latest
legal	awesome	useful	global
happy	graphic	different	graphic
financial	rocky	happy	golden
electronic	indoor	important	virtual
<i>Average rank for 10 most frequent unambiguous adjectives (N=5000)</i>			
203	530	158	359
<i>Average rank for the 100 most frequent unambiguous adjectives (N=5000)</i>			
1483	2176	1563	1932

4. Conclusion

We have begun an exploration of the different language uses in English language indexes, comparing certain grammatical and semantic types in four different collections: the general web pages indexed by Exalead, the edited text found in Wikipedia, the text found new or pointing to images indexed by YahooImages, and the manually supplied tags added by Flickr image suppliers. All of these language populations are uncontrolled, unrestricted text. We find that there are discernible differences. Flickr users prefer using nouns to adjectives, though color terms are popular. Since Flickr attracts many people storing their vacation photos, we find many pictures indexed with place names, activity names, and, maybe surprisingly, food. Abstract nouns and longer words are found less than in written text. The automatically assigned words of YahooImages tend to following the same patterns as manually assigned tags in Flickr, though the language model is also more similar to

that found in Wikipedia.

This study continues our exploration of the differences between the language used to describe images and that used in ordinary written text. There are differences, but there is still a lot of work to be done to circumscribe these differences.

5. Acknowledgements

This research is being partially funded by a grant from the Fondation Jean-Luc Lagardère, and by the French DGE competitiveness pole project POPS (07.2.93.0462).

6. References

- W. Bruce Croft, John Lafferty. Language Modeling for Information Retrieval. Kluwer, 2003
- R. Datta, D. Joshi, J. Li, and J. Z. Wang, Image retrieval: Ideas, influences, and trends of the new age," ACM Computing Surveys, 40, 2008
- Gregory Grefenstette. Conquering Language: Using NLP on a Massive Scale to Build High Dimensional Language Models from the Web. CICLing 2007: 35-49
- Gregory Grefenstette, Guillaume Pitel: Image Specific Language Model: Comparing Language Models from Two Independent Distributions from Flickr and the Web. CICLing 2008, Haifa, Feb 2008
- Allan Hanbury. Analysis of Keywords used in Image Understanding Tasks. OntoImage International Workshop, Genova, Italy; 2006
- M. Junghoefler, Bradley, M. M., Elbert, T. R., & Lang, P. J. Fleeting images: A new look at early emotion discrimination. Psychophysiology, 38, 175-178., 2001
- Jia Li, Shih-Fu Chang, Michael Lesk, Rainer Lienhart, Jiebo Luo, Arnold W. M. Smeulders: New challenges in multimedia research for the increasingly connected and fast growing digital society. Multimedia Information Retrieval 2007: 3-10
- C. Marlow, Naaman, M., Boyd, D., and Davis, M.. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, Odense, Denmark, Aug. HYPERTEXT '06. ACM, New York, NY, 31-40. 2006
- Guillaume Pitel, Christophe Millet, Gregory Grefenstette: Deriving a Priori Co-occurrence Probability Estimates for Object Recognition from Social Networks and Text Processing. ISVC (2) 2007: 509-518
- N. Vasconcelos: From Pixels to Semantic Spaces: Advances in Content-Based Image Retrieval, Computer, 40:7, 20-26, 2007

Semantic-based Expansion of Image Retrieval Queries

Taha Osman¹, Dhavalkumar Thakker¹, Gerald Schaefer², Thomas Geslin³, Philippe Kiffer³

¹ School of Computing & Informatics, Nottingham Trent University
Nottingham, NG11 8NS, UK

² School of Engineering & Applied Science, Aston University
Aston Triangle, Birmingham B4 7ET, UK

³ Ecole Navale

Naval Academy of France, BP 600, 29240 Brest, France

E-mail: ¹{taha.osman,dhavalkumar.thakker}@ntu.ac.uk, ²g.schaefer@aston.ac.uk, ³contact: claramunt@ecole-navale.fr

Abstract

The proliferation of digital media has led to a huge interest in classifying and indexing media objects for generic search and usage. In particular, we are witnessing colossal growth in digital image repositories that is difficult to navigate using free-text search mechanisms that often return inaccurate matches as they in principle rely on statistical analysis of query keyword recurrence in the image annotation or surrounding text. In this paper we present a semantically-enabled image annotation and retrieval engine that relies on methodically structured ontologies for image annotation, thus allowing for more intelligent reasoning about the image content and subsequently obtaining a more accurate set of results and a richer set of alternatives matchmaking the original query. We show how well-researched and designed domain ontology contributes to the implicit expansion of user queries as well as presenting our initial thoughts on exploiting lexical databases for explicit semantic-based query expansion.

1. Introduction

The last few years have witnessed an unprecedented interest in digital media and subsequently colossal growth of public and commercial digital media repositories (audio, images, and video). Retrieving relevant media from these ever-increasing repositories is an impossible task for the user without the aid of search tools. Most public image retrieval engines rely on analysing the text accompanying the image to matchmake it with the user query. Various optimisations were developed including the use of weighting systems where for instance higher regard can be given to the proximity of the keyword to the image location, or advanced text analysis techniques that use term weighting method, which relies on the proximity between the anchor to an image and each word in an HTML file (Fuji 2005). Despite the optimisation efforts, these search techniques remain hampered by the fact that they rely on free-text search that, while cost-effective to perform, can return irrelevant results as it primarily relies on the recurrence of exact words in the text accompanying the image. The inaccuracy of the results increases with the complexity of the query. For instance, while performing this research we used the Yahoo™ search engine to look for images of the football player Zico returns some good pictures of the player, mixed with photos of cute dogs (as apparently Zico is also a popular name for pet dogs), but if we add the action of scoring to the search text, this seems to completely confuse the Yahoo search engine and only

one picture of Zico is returned, in which he is standing still!

Any significant contribution to the accuracy of matchmaking results can be achieved only if the search engine can “comprehend” the meaning of the data that describes the stored images, for instance, if the search engine can understand that scoring is an act associated with sport activities performed by humans. Semantic annotation techniques have gained wide popularity in associating plain data with “structured” concepts that software programs can reason about (Wang 2006).

In our recent publication (Osman 2007) we present comprehensive coverage of our integrative framework for semantic-based image retrieval, but this contribution focuses in particular on the query expansion aspects of our work. We claim that shrewd analysis of the application domain characteristics, coupled with a subsequently well-designed ontology can significantly contribute to the user query expansion process via direct term replacement or by modifying the query’s class structure. We also present our initial research into using lexical databases to analyze free-entry queries in our effort to make them compatible with the requirements of our semantic search engine.

The paper begins with an overview of the Semantic web technologies. In section 3 we review the case study that was the motivation for this work. Section 4 overviews the engineering of the ontology, and the annotation and retrieval mechanism. Section 5 details

our strategy for query expansion. We present our conclusions section 6.

2. Semantic-based image retrieval

The fundamental premise of the semantic web is to extend the Web's current human-oriented interface to a format that is comprehensible to software programmes. Naturally this requires a standardised and rich knowledge representation scheme or Ontology.

This comprehensive representation of knowledge from a particular domain allows reasoning software to make sense of domain-related entities (images, documents, services, etc.) and aid in the process of their retrieval and use.

Applied to image retrieval, the use of Semantic technologies can significantly improve the computer's understanding of the image objects and their interactions by providing a machine-understandable conceptualisation of the various domains that the image represents. This conceptualisation integrates concepts and inter-entity relations from different domains, such as Sport, People and Photography. In relation to the "Zico scoring goal" query discussed in the introduction, a semantic search engine can infer that Zico a *person* and thus can take *actions*, and because he is a *footballer*, the action can be *scoring* a goal, and that he used to be a player of the *Brazil national team*, who lost the *World Cup* final in 1986, etc.

3. Case study for semantic image retrieval

An opportunity to experiment with our research findings in semantic-based search technology was gratefully provided by PA Photos™. PA Photos is a Nottingham-based company which is part of the Press Association Photo Group Company (PA Photos 2007). As well as owning a huge image database in excess of 4 million annotated images which date back to the early 1900's, the company processes a colossal amount of images each day from varying events ranging from sport to politics and entertainment. The company also receives annotated images from a number of partners that rely on a different photo indexing schema.

More significantly, initial investigation has proven that the accuracy of the results sets matching the user queries do not measure up to the rich repository of photos in the company's library.

The objective of the case study is two-fold is to investigate the use of semantic technology to build a classification and indexing system that critically unifies the annotation infrastructure for all the sources of incoming stream of photos, and subsequently conduct a feasibility study aiming to improve the end-user experience of their images search engine. At the moment PA Photos search engine relies on Free-Text search to return a set of images matching the user requests. Therefore the returned results naturally can go off-tangent if the search keywords do not exactly recur in the photo annotations. A significant improvement

can result from semantically enabling the photo search engine. Semantic-based image search will ultimately enable the search engine software to understand the "concept" or "meaning" of the user request and hence return more accurate results (images) and a richer set of alternatives.

It is important here to comment about the dynamics of the retrieval process for this case study as it represents an important and wide-spread class of application areas where there is a commercial opportunity for exploiting semantic technologies:

1. The images in the repository have not been extracted from the web. Consequently the extensive research into using the surrounding text and information in the HTML document in improving the quality of the annotation such as in Maina (2005) is irrelevant.
2. A significant sector of this market relies on fast relay of images to customers. Consequently this confines advanced but time-consuming image analysis techniques (Lam 2006) to off-line aid with the annotation of caption-poor images.

4. Ontology development

4.1 Domain Analysis

Our domain analysis started from an advanced point as we had access to the photo agency's current classification system. Hence, we adopted a top-down approach to ontology construction that starts by integrating the existing classification with published evidence of more inclusive public taxonomies Roach (2008). At the upper level, two ontological trees were identified; the first captures knowledge about the event (objects and their relationships) in the image, and the second is a simple upper class that characterises the image attributes (frame, size, creation date, etc.), which is extensible in view of future utilisation of content-recognition techniques.

At the initial stages of the research, we decided to limit our domain of investigation to sport-related images. A bottom-up approach was used to populate the lower tiers of the ontology class structure by examining the free-text and non-semantic caption accompanying a sample set of sport images. Domain terms were acquired from approximately 65k image captions. The terms were purged of redundancies and verified against publicly available related taxonomies such as the media classification taxonomy detailed in (Roach 2008). An added benefit of this approach is that it allows existing annotations to be seamlessly parsed and integrated into the semantic annotation.

Wherever advantageous, we integrated external ontologies (such as the aktors ontology in (AKT (2006))) into our knowledge representation. However, bearing in mind the responsiveness requirements of on-line retrieval applications, we applied caching methods to

localise the access in order to reduce its time overhead.

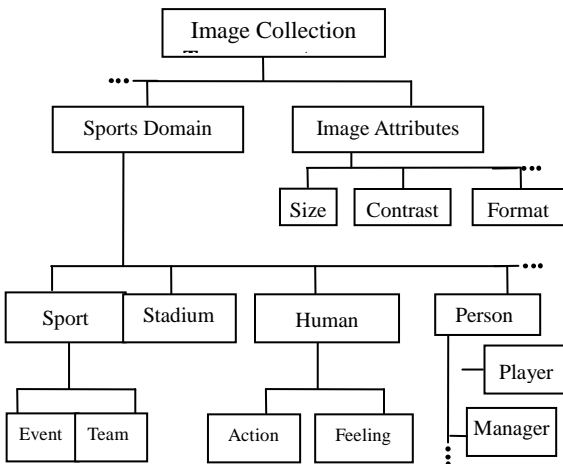


Figure 1: Subset of the ontology tree.

4.2 Normalisation: reducing the redundancy

The objective of normalisation is to reduce redundancy. In ontology design, redundancy is often caused by temporal characteristic that can generate redundant information and negatively affect the performance of the reasoning process.

For instance, direct adoption of the ontology description in Figure 2 below will result in creating new team each season, which is rather inefficient as the team should be a non-temporal class regardless of the varying player's membership or tournament participation every season. Hence, Arsenal or Glasgow Rangers Football clubs need to remain abstract entities.

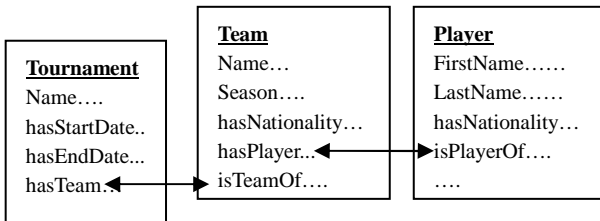


Figure 2: non-normalised ontology design

Our approach was to introduce an intermediary temporal membership concept that serves as an indispensable link between teams and players, as well as between teams and tournaments as illustrated in Figure 3 below.

The temporal instances from the Membership class link instances from two perpetual classes as follows:

- memberEntity links to a person (Player, Manager, Supporter, Photographer, etc.);

- isMemberOf refers to the organisation (Club, Press Association, Company, etc.);
- fromPeriod and toPeriod depict membership temporal properties

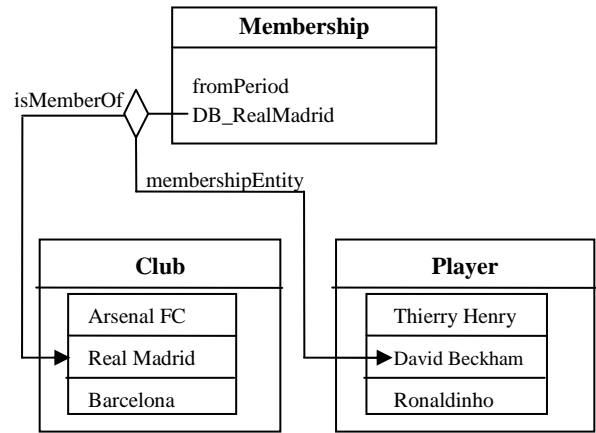


Figure 3: Normalisation using Membership class

5. Overview of the annotation and retrieval processes

The Web Ontology Language (OWL 2004) was adopted to annotate image captions and Jena (Carroll 2004) java API was used to build the annotation portal to the constructed ontology.

Taking into account the dynamic motion nature of the sport domain, our research concluded that a variation of the sentence structure suggested in (Hollink 2003) is best suited to design our annotation template. We opted for an “Actor – Action/Emotion – Object” structure that will allow the natural annotation of motion or emotion-type relationships such as “Beckham – Smiles – null”, or “Gerrard – Tackles – Henry”, with a view of more seamless utilisation of NLP techniques (Chen 1995) for query expansion.

The image retrieval user interface is illustrated in Figure 4. The search query can include sentence-based relational terms (Actor-Emotion/Action-Object) and/or key domain terms (such as tournament and team). In case multiple terms were selected for the query, the user needs to specify which term represents the main search preference (criterion).

For instance, in Figure 4 the relational term (Gerrard Tackles Rooney) is the primary search term and team Liverpool is the secondary search term. The preference setting is used to fine-tune the ranking of retrieved images.

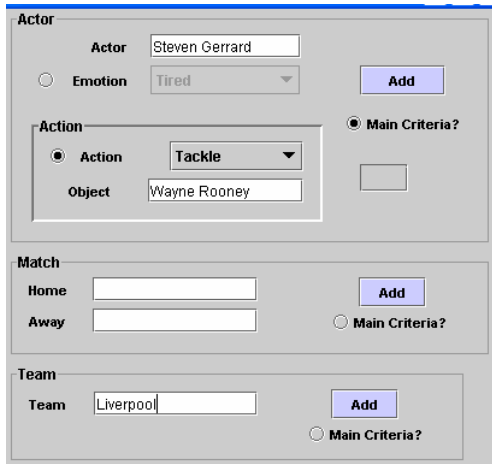


Figure 4: Snapshot of the retrieval interface

The semantic reasoning engine applies our matchmaking algorithm at two phases: The first phase retrieves images with annotations matching all concepts in the query, while in the second phase further matchmaking is performed to improve the ranking of the retrieved images in response to user preferences. Our reasoning engine uses a variation of the nearest neighbour matchmaking algorithm (Osman 2006) to serve both the semantic retrieval and the ranking phases. The algorithm continues traversing back to the upper class of the ontology and matching instances until there are no super classes in the class hierarchy, i.e. the leaf node for the tree is reached, giving degree of match equal to 0. The aggregate degree of match ($ADoM$) is calculated according to the following equation:

$$ADoM = \sum_{i=1}^n W_i \times \frac{MN_i}{GN_i}, \forall W_i \in [0,1]$$

Equation 1: Aggregate Degree of Match

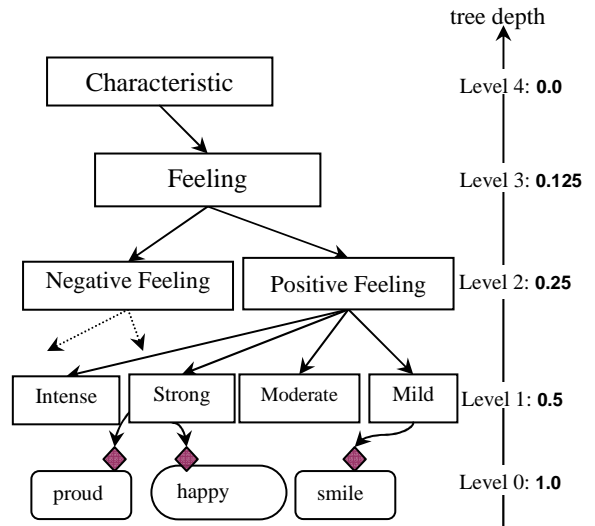
Where the MN is the total number of matching nodes in the selected traversal path, and GN the total number of nodes in the selected traversal path for a particular matching criterion (i). Each criterion is scaled with the importance factor W according to the user preferences.

The example below illustrates the operation of the algorithm for a single criterion only where the query is: Object- hasCharacteristic-happy, and image1 and image2 are annotated with Object-hasCharacteristic-happy and Object-hasCharacteristic-smile respectively, the DoM for image1 is 1 as the instances match to the level of the leaf node (Figure 5). However, for image2 instances match to the level of Positive Feeling- Mild class and is one layer lower than the leaf node giving $DoM = 0.5$.

Figure 5: Traversing the Ontology Tree

6. Strategy for Query Expansion

Lately query expansion (QE) techniques have gained a lot of attention in attempting to improve the recall of document and media queries. Query expansion is traditionally considered as a process of supplementing



a query with additional terms as the assumption is that the initial query as provided by the user may be an inadequate representation of the user's information needs (Gaihua 2005) (Croft 1996).

Taking into account the domain knowledge hardwired into the ontologies, semantic-based query expansion techniques can broadly be classified in two categories: implicit and explicit. Implicit query expansion can be considered as a by-product of well-researched and designed domain ontology.

The "Actor-Action/Emotion-Object" semantic format allows to naturally employing the ontology to find related terms via simple equivalence relations as that of equating the action of *smiling* to the emotion of *happiness*. Taking into account the limited vocabulary of the *sport* domain, in consultation with the domain experts, we decided against the automatic expansion of directly related terms from a lexical public database such as WordNet (Fellbaum 1998). Our initial experiments have shown that while that expansion improved image recall, the accuracy of returned results suffered significantly particularly for complex queries where partial replacement of terms might invalidate the semantics of the query.

Using our ontology structure we are also able to expand queries implicitly by analyzing more complex relations as in inferring that *Liverpool* is a possible replacement for *Chelsea* as both are *Teams* playing *Football* sport in the *Premier League* in *England*. Moreover, we are able to scale the relatedness of each term in the query tree according to the importance weighting set by the user/domain manager as explained in the previous section.

Explicit query expansion involves direct replacement of terms in the user query with terms that were identified as identical by the knowledge domain

administrator or the end user. These replacement terms are not part of the ontology infrastructure, but are kept in a separate synonym dictionary that contains one-to-many (USE_FOR) relations between the ontology term and the possible synonyms. For instance, the domain administrator might use the ontology term “Manchester United” to replace the popular term “Man UTD”. Similarly, users are allowed to cache (USE_FOR) terms on the client-side for exclusive expansion of their queries. The domain administrator has access to the most popular cached nicknames/synonyms and can choose to enter them into the main synonym dictionary.

We considered adding synonyms to the ontology using OWL’s owl:sameAs property, but decided against it primarily because of the performance penalties in processing RDF data as opposed to simple text strings. We also think that from a pure semantic engineering point of view, nicknames such as “Man UTD” should not exist as an RDF individual.

Finally, we started considering using NLP techniques to attempt to translate free-entry queries that are not constructed using our domain-tailored retrieval interface (see Figure 4) into our “Actor-Action/Emotion-Object” semantic format to allow for semantic reasoning.

At the time of writing this paper we succeeded in utilising WordNet lexical database primarily in identifying verbs that might be candidate for the “Action/Emotion” central part of our annotation format. Subsequently, the left part to the verb is further analyzed as an “actor” candidate, and the right as an “object” candidate, applying our spelling checker and synonym replacement where appropriate. For instance the free-entry: “Man Utd’s Wayne Rooney tackles the French player Zizou” is analyzed as follows:

Man Utd’s Wayne Rooney	tackles	the French player Zizou.
Subject part	Verb	Object part

Man Utd	s	Wayne Rooney	tackles
Manchester United	#	Wayne Rooney	Tackle

the	French	player	Zizou
###	adj		Zinedine Zidane

Hence, the sentence analyzer infers the request below which can be now fired at our semantic image retrieval engine:

Actor	Wayne Rooney
Action	Tackle
Object	Zinedine Zidane
Team	Manchester United

7. Conclusion

In this paper we presented a comprehensive solution for

image retrieval applications that takes full advantage of advances in semantic web technologies to coherently implement the annotation, retrieval and query expansion components of the integrative framework.

The first stage of the development was producing ontologies that conceptualise the objects and their relations in the selected domain. We methodically verified the consistency of our ontology, optimised its coverage, and performed normalisation methods to rid of concept redundancies. Our annotation approach was based on a variation of the “sentence” structure to obtain the semantic-relational capacity for conceptualising the dynamic motion nature of the targeted sport domain. This careful analysis of the domain features allowed us to hardwire application domain knowledge into the ontology and hence implicitly perform query expansion either by simple replacement of equivalent terms or by traversing the ontology tree to modify more complex queries.

The retrieval algorithm is based on a variation of the nearest-neighbour search technique for traversing the ontology tree and can accommodate complex, relationship-driven user queries. The algorithm also provides for user-defined weightings to improve the ranking of the returned images and was extended to embrace query expansion technology in a bid to improve the quality of the recall.

We also presented our initial research into using lexical databases to analyze free-entry queries in our effort to make them compatible with the entry requirements of our semantic search engine.

8. References

- AKT (2006). Advanced Knowledge Technologies. <http://www.aktors.org/ontology/portal#>
- Carroll, J. *et al* (2004). “Jena: implementing the semantic web recommendations”, Proceedings of the 13th international World Wide Web conference, New York, USA, ACM Press, pp. 74-83.
- Chen, H. (1995). “Machine Learning for information retrieval: Neural networks, symbolic learning and genetic algorithms”, Journal of the American Society for Information Science and Technology, 46(3), April 1995, pp. 194-216.
- Fellbaum, C. (1998). “WordNet: An Electronic Lexical Database and Some of its Applications”. MIT Press.
- Fuji A., Ishikawa, T. (2005). “Toward the Automatic Compilation of Multimedia Encyclopaedias: Association Images with Term Descriptions on the Web”, In Proceedings of the 2005 International Conference on Web Intelligence – WIC05, Compiègne, France, September 19-22, 2005, pp. 536-542.
- Gaihua, F. *et al* (2005). “Ontology-based Spatial Query Expansion in Information Retrieval”. Lecture Notes in Computer Science, Vol. 3761/2005, pp. 1466-1482.
- Hollink, L. *et al* (2003). “Semantic annotation of image collections. In Workshop on Knowledge Markup and

- Semantic Annotation”, KCAP’03, Florida, USA, 2003.
- Lam, T., Singh, R. (2006). "Semantically Relevant Image Retrieval by Combining Image and Linguistic Analysis", Proc. International Symposium on Visual Computing (ISVC), Lecture Notes in Computer Science Vol. 4292, pp. 1686 - 1695, Springer Verlag, 2006
- Maina E.W. *et al* (2005). "Semantic Image Retrieval Based On Ontology and Relevance Model: A Preliminary Study", Digital Engineering Workshop, Tokyo, Japan, 24-25 February, 2005, pp. 331-339.
- Osman, T. *et al* (2007). "An integrative Framework for Image Annotation and Retrieval. In the proceedings of the IEEE International Conference on Web Intelligence WI’07.
- Osman, T., Thakker, D. (2006), "Semantic-Driven Matchmaking of Web Services Using Case-Based Reasoning" In proceedings of IEEE International Conference on Web Services (ICWS’06), Chicago, USA, September 2006. pp. 29-36.
- OWL (2004). OWL Web Ontology Language Overview. <http://www.w3.org/TR/owl-features>.
- PA Photos (2007). <http://www.paphotos.com/>
- Roach, M. *et al.* (2002), N. Evens, L. Xu, F. Stentiford, "Recent Trends in Video Analysis: A Taxonomy of Video Classification Problems", Internet and Multimedia Systems and Applications, Kauai, Hawaii, USA, 2002, pp.348-353.
- Wang, H., Liu, S. (2006). "Does ontology help in image retrieval? A comparison between keyword, text ontology and multi-modality ontology approaches", Proceedings of the 14th annual ACM international conference on Multimedia, Hawaii, USA, pp. 109 – 112
- Xu, J., Croft, W. (1996). "Query expansion using local and global document analysis". Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp 4-11.

Combining Medical Domain Ontological Knowledge and Low-level Image Features for Multimedia Indexing

Dina Demner-Fushman, Sameer K. Antani, Matthew Simpson, George R. Thoma

Lister Hill National Center for Biomedical Communications,
National Library of Medicine, NIH, Bethesda, MD
{ ddemner, santani, simpsonmatt, gthoma } @mail.nih.gov

Abstract

Biomedical images are invaluable in establishing diagnosis, acquiring technical skills, and implementing best practices in many areas of medicine. At present, images needed for instructional purposes or in support of clinical decisions appear in specialized databases and in biomedical articles, and are therefore not easily accessible. Our goal is to automatically annotate images extracted from scientific publications with respect to their usefulness for clinical decision support and instructional purposes, and project the annotations onto images stored in databases by linking images through content-based image similarity. This paper presents an overview of our approach to automatic image indexing, content-based image analysis, and the results of a pilot evaluation of an automatic indexing method based on biomedical terms extracted from snippets of text pertaining to images appearing in scientific biomedical articles.

1. Introduction

Essential information is often conveyed in illustrations in biomedical publications. These images can be used to illuminate document summaries and answers to clinical questions, to enrich large image collections with textual information from articles, and for instructional purposes. The problem however is to automatically determine which of the images in an article will best serve each of the aforementioned purposes. Our approach to automatic image indexing is to describe (or annotate) an image at three levels of granularity:

- **coarse**, which addresses
 - image modality,
 - relation to a specific clinical task (image utility),
 - body location;
- **medium**, which provides a more detailed description of the image using existing biomedical domain ontologies;
- **specific**, which provides very detailed descriptions of clinical entities and events in an image using terms that are not included in existing ontologies and often are familiar only to clinicians specializing in a narrow area of medicine.

In this paper, we present a pilot evaluation of medium-level indexing that can be achieved by automatically extracting biomedical terms currently available in the largest biomedical domain ontology, the Unified Medical Language System[®] (UMLS[®]) Metathesaurus, from snippets of text pertaining to images in scientific biomedical articles (image captions and relevant discussion in the text). We also provide an overview of our research in coarse- and specific-level image indexing and content-based image analysis.

2. Background

In our previous exploration of coarse automatic indexing of images by modality (color image, gray-scale image,

graph, graphic illustration, etc.) and image utility (suggested by the Evidence Based Medicine paradigm six elements of a clinical scenario that an image might illustrate), we combined image and textual features in a supervised machine learning approach. Textual features were obtained from the captions to the images and paragraphs of text containing discussion (“mentions”) of these images. The text and the images were automatically extracted from the HTML-formatted articles. Text was represented as a bag-of-words or as a set of terms obtained by mapping these captions and mentions to the UMLS Metathesaurus. Texture and color features were computed on the entire image without applying any image segmentation techniques.

Texture features were computed as a 3-level discrete 2-D Daubechies’ wavelet transform. The four most dominant colors were computed in the perceptually uniform CIE LUV color space and proved most effective. At this coarse level of granularity, a multi-class SVM classifier trained on a bag-of-words representation of image captions performed better in determining image modality ($84.3\% \pm 2.6\%$ accuracy) than when trained on a combination of textual and image features or features reduced to the domain specific vocabulary. For image utility, however, the combination of image and textual features was better than any single-source feature set achieving $76.6\% \pm 4.2\%$ accuracy (Demner-Fushman et al., 2007).

Often in biomedical publications, several images are combined into a multi-panel figure. This requires sub-figure separation for image analysis to determine image modality. We therefore developed a two-phase algorithm to detect and separate figure panels using cues from caption text analysis, horizontal and vertical profiles and panel edge information (Antani et al., 2008). Further analysis on each image panel revealed its coarse modality. For instance, using color histogram profiles we could determine with sufficient precision if an image is a color

image, an illustration/drawing, or a radiographic image (CT, MRI, x-ray, sonogram, etc.). Detecting image modalities is useful in further image analysis and sub-categorization. Our efforts in this area resulted in development of a method for detecting text overlays on images, arrows, and other content valuable for indexing images by visual content and correlated text description (Antani et al., 2008).

2.1 Prior Work in Content-Based Image Retrieval

Our image analysis and image indexing work stems from an ongoing long-term research and development effort into image understanding and content-based image retrieval (CBIR) of biomedical images. We have worked with a large collection of digitized x-ray images of the spine derived from a nationwide health survey to develop image segmentation techniques for extraction of vertebral shape information important to researchers of osteoarthritis and musculoskeletal diseases. Whole and partial shape similarity techniques, multiple object similarity, multidimensional data indexing, relevance feedback, and Web-based frameworks for CBIR have been explored (Hsu et al., 2007).

Subsequently the research has been expanded into localization and similarity matching of pre-cancerous lesions in the uterine cervix on a data set acquired by the National Cancer Institute (NCI) from a multi-year longitudinal study. For this dataset color, texture, and location methods were studied to enable CBIR of several types of regions of interest (Xue et al., 2007). As both data sets have free-text medical records corresponding to the images, we have explored combined text and image retrieval on this data.

Finally, we have also explored automatic coarse-level image labeling and classification on the ImageCLEF 2005 data set using Semantic Error-Correcting Output Codes (SECC) and achieved an overall error rate of 18.7 using 9,000 training images and 1,000 test images (Yao et al., 2006).

Coarse-level image indexing is not sufficient to describe an image taken from a publication beyond achieving retrieval of a particular modality, utility, and location, for example, *ultrasound images for diagnosis of heart conditions*. We hypothesize that medium-level image annotation will facilitate finding images to illustrate summaries and answers to clinical questions, for example, about *echocardiographic finding of mitral annular calcification*. Specific-level indexing will be required to answer detailed questions, such as *What is the efficacy of thick acellular human dermis grafts for posterior and middle lamellae reconstruction?*

3. Methods

To automatically achieve medium-level indexing we extracted the image captions and mentions from the article text and processed the text using MetaMap, a tool that maps biomedical text to the UMLS (Aronson, 2001).

The indexing terms were extracted from the MetaMap machine output, which provides comprehensive information about the mappings of phrases found in the text to the UMLS concepts. The following information was retained: the concept unique identifier (CUI) and semantic type, the preferred UMLS name for the concept, and the offset and length of the substring that was mapped to the concept.

To enable content experts to evaluate the quality of the extracted indexing terms we developed a Web-based evaluation and annotation interface (see Figures 1 and 2). This interface displays an image, bibliographic information about the article from which the image was extracted, and two tabs for annotation and evaluation. The first tab shown in Figure 1 is used for coarse-level image annotation through selecting pre-defined indexing terms for modality, utility and body location. The second tab (Figure 2) serves two purposes:

1. Evaluation of the automatically extracted indexing terms for medium-level indexing;
2. Manual annotation of the image with specific terms, more fine-grained than currently available in the UMLS (specific-level indexing), such as *thick acellular human dermis graft*. Parts of this term can be mapped to the UMLS, but even the closest existing term *Acellular Dermal Replacement* cannot be mapped to the specific term using existing tools.

The purpose of the manual annotation is to identify such missing terms and establish their ontological relations. The results of manual annotation will be used for development and evaluation of automatic indexing methods on all three levels of granularity.

The indexing terms and ontological information extracted from the MetaMap output (Figure 2 top) were evaluated on two axes:

1. Usefulness in image indexing, evaluated on a binary scale.
2. Relevance to the image, evaluated on a five-point scale, ranging from an *exact match* to *unrelated*.

An identified term might not be useful for indexing if it is too broad, too narrow, or unrelated to the image. An unrelated term might be extracted for two reasons:

1. A term might be extracted from the caption text verbatim, but the senses of the term available in the UMLS are not relevant to the image. For example, the string *apex* identified in the caption *Thrombus in left ventricular apex* maps through synonymy to the UMLS concepts:
 - APEX1 gene
 - APEX1 protein, human
 - Highest

The UMLS Metathesaurus does not contain the term *ventricular apex*; and mapping to the correct sense *Cardiac apex* is not possible using strict matching, because the set of synonyms for the *Cardiac apex* concept does not include the term *apex*.

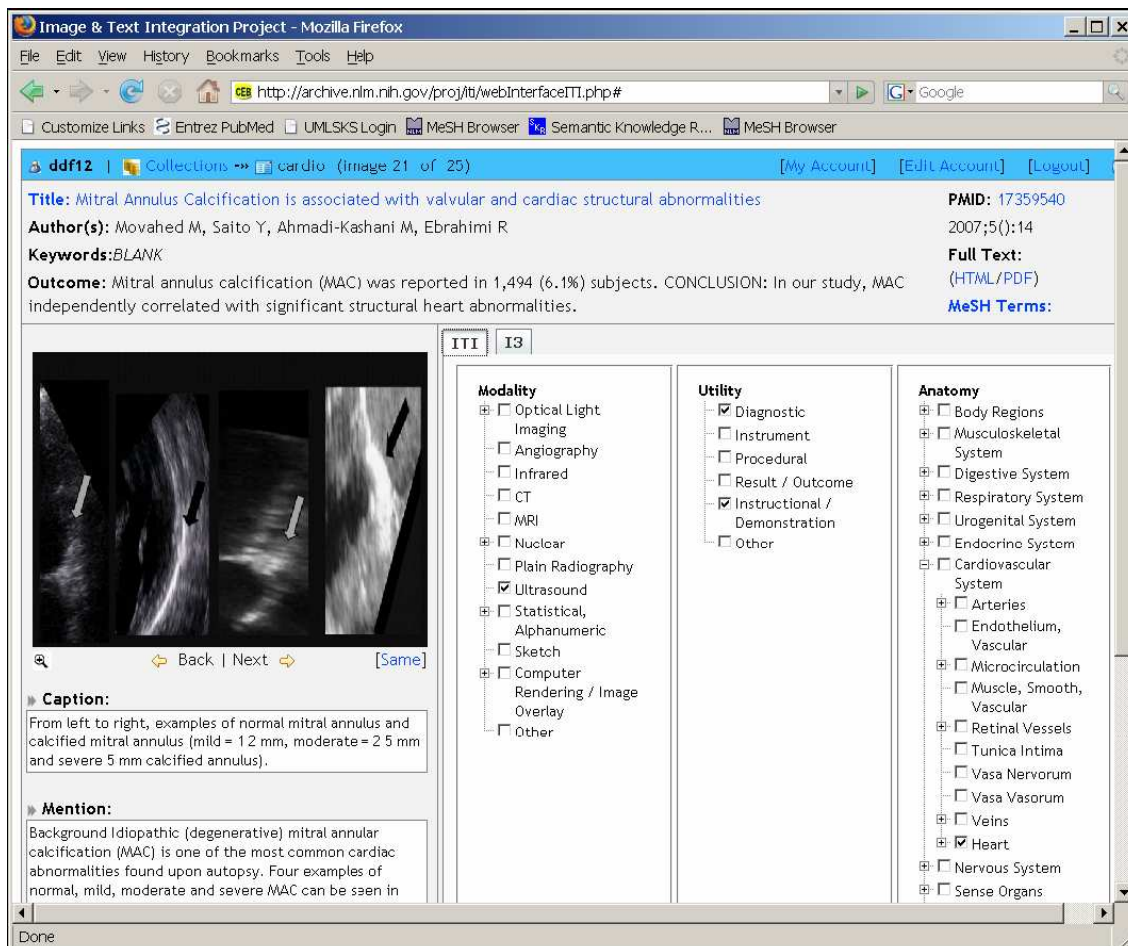


Figure 2: A Web-based application for image indexing annotation and evaluation. Coarse-level annotation categories.

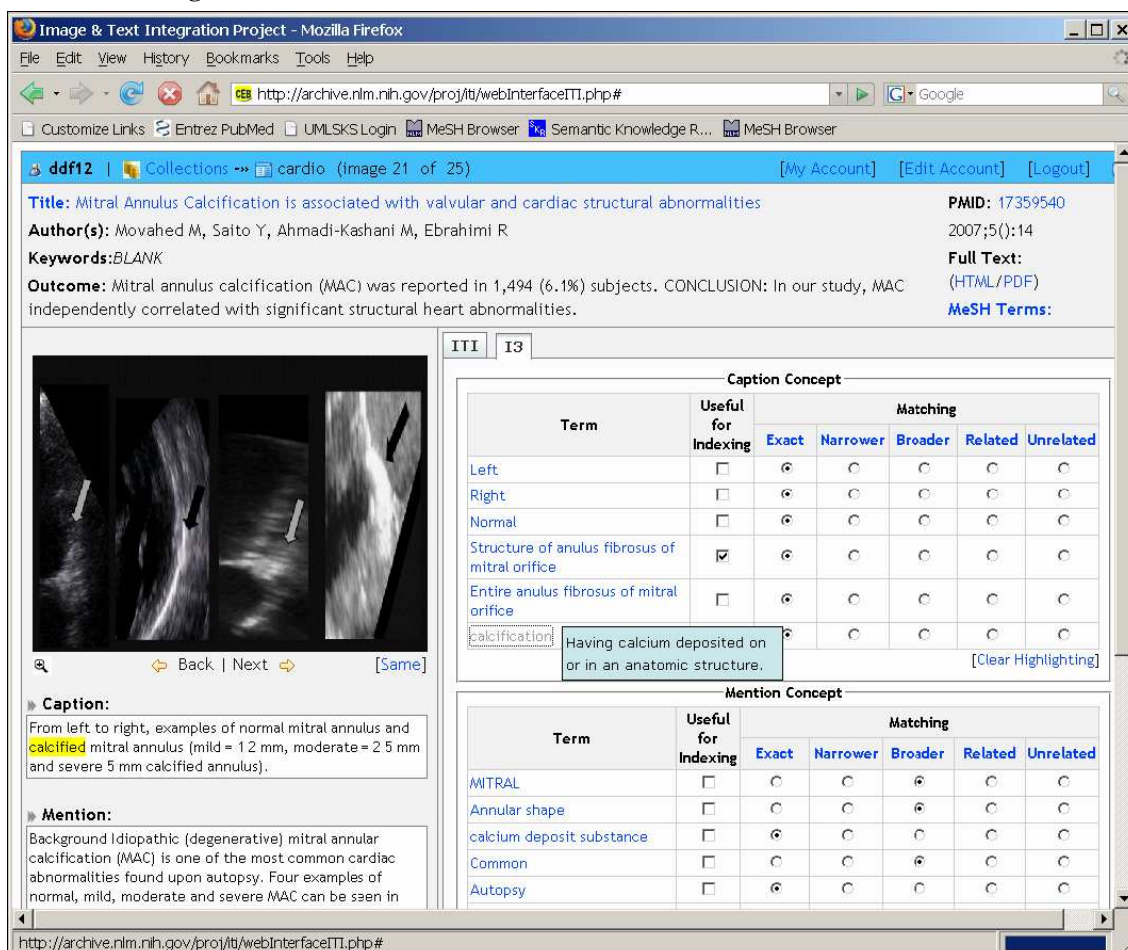


Figure 1: A Web-based application for image indexing annotation and evaluation. Medium-level indexing evaluation.

2. A substring identified in the text could be matched to a wrong term in the UMLS Metathesaurus because it is an acronym or abbreviation. For example, the term *LV* identified in the caption an initial increase of *LV filling pressure* is synonymous with:

- Latvia
- Leucovorin Calcium
- Liposome Vesicle

The UMLS Metathesaurus does not contain the expansion of *LV* to *left ventricular* expected in the context of cardiovascular imaging. The assumption that only this sense of the term is expected in the context of cardiovascular imaging is based on the observation that the term is not expanded anywhere in the paper containing the image.

Our interface tool assists the evaluators in determining the sense of the extracted terms through the UMLS definitions which are displayed by positioning the computer mouse over the suggested index term. The tool retrieves the UMLS definitions using the extracted unique concept identifiers. Assistance for determining the origin of an extracted term is provided through highlighting the substring that was mapped to a term in the caption or mention text upon clicking on the suggested index term.

The evaluation interface was used by five physicians and one medical imaging specialist who manually assigned missing specific terms, and evaluated the quality of medium-level indexing terms. The indexing terms were automatically extracted from captions and descriptions of 50 images randomly selected for each evaluator from all images published in *BMC Annals of Facial and Plastic Surgery* and *European Journal of Cardiovascular Imaging* during 2006 and 2007. Their judgments were analyzed to answer the following questions:

1. Do captions and mentions of the image in the text provide information beyond indexing terms assigned by NLM indexers to the papers containing those images?
2. Is the extracted text sufficient for image annotation?
3. Is our extraction method satisfactory?

The first question was answered by intersecting the extracted terms evaluated as useful for imaging with the indexing terms assigned to the papers by NLM indexers and extracted from the bibliographic citations to the papers. These citations in XML format were retrieved using PubMed/MEDLINE®.

The second question was answered by intersecting the additionally assigned terms with the extracted text and with the full-text paper.

The extraction method was evaluated using recall and precision computed for each evaluator as follows: The desired index terms D for the images are the set of extracted terms evaluated as useful for indexing

combined with the indexing terms added by the evaluator, A is the set of all suggested indexing terms, and within A there is a set of terms evaluated as useful for indexing C. Precision P and recall R are:

$$P = |C|/|A|$$

$$R = |C|/|D|$$

Precision and recall were computed for each evaluator, and then averaged.

4. Results

The six evaluators scored 4, 006 concepts (3, 281 of which were unique) pertaining to 186 unique images extracted from 109 papers. Table 1 presents the average numbers of concepts per image evaluated and found useful for indexing by each evaluator. The majority of the terms rated useful for indexing were also rated as an exact match.

Table 1: Average number of concepts per image.

Evaluators trained in medical informatics are marked with an asterisk.

Specialty	Indexing Terms		
	evaluated	useful	%useful
family physician*	19.26	2.38	12.4%
cardiologist*	17.80	2.02	11.4%
plastic surgeon*	17.89	1.80	10.1%
internist*	17.55	2.18	12.4%
general surgeon	19.98	1.50	7.5%
medical imaging	14.46	1.40	9.9%
Mean ± CI	17.83±2.0	1.89±0.4	10.6±2.0%

The 349 exact matches constitute 77.4% of the terms marked as useful for indexing. The remaining 102 selected indexing terms were rated primarily as being broader than an exact description of the image would warrant.

4.1 Indexing terms assigned to the article and image annotation

Overall, the evaluators rated 451 extracted terms as useful for indexing and submitted 255 additional indexing terms.

Table 2: Match between indexing terms assigned to images and papers.

Evaluators trained in medical informatics are marked with an asterisk.

Specialty	MeSH Terms		
	extracted	added	%used
family physician*	33.0%	34.9%	11.5%
cardiologist*	39.8%	48.7%	20.5%
plastic surgeon*	46.9%	41.2%	11.1%
internist*	25.0%	25.7%	11.7%
general surgeon	33.3%	---	7.1%
medical imaging	28.8%	---	5.3%
Mean ± CI (%)	34.5±8.2	25.1±21.9	11.2±5.5

Table 2 presents the percentages of terms assigned by the evaluators that match terms assigned by NLM indexers (MeSH terms) to the papers containing the images. In

addition, the %used column of the table shows the proportion of the MeSH terms assigned to the paper that were deemed useful in annotating images.

4.2 Locating additional terms in the text

For three of the 255 indexing terms added by the evaluators no image-related text was extracted. Of the remaining 252 added terms, 75 were extracted verbatim from the caption text and 11 from the discussion of the image in the text. Another 139 added terms were generated using captions and mentions through:

- extracting strings with gaps, for example, extracting *Preoperative photograph* from *Preoperative and postoperative photographs*;
- paraphrasing, for example, deriving *elderly* from *89-year old*;
- summarizing, for example, the following mention of the image: *a mobile, left-sided, nasal dorsal implant with tip ptosis, erythema, and swelling of the left nasal vestibule as implantation complications*;
- generalizing based on the figure and the caption, for example, *ultrasound*; *surgical method*; or *transthoracic echocardiography*.

The remaining 27 terms were found in the paper title, abstract, and MeSH terms assigned to the paper. Of the 255 additionally assigned terms 103 were subsequently mapped to the UMLS concepts.

4.3 Extraction accuracy

The design of the extraction evaluation was recall oriented. All extracted terms were given to the evaluators without any filtering to have enough training examples for learning term selection in the future. Recall and precision achieved by this baseline extraction method are shown in Table 3.

Table 3: Evaluation of the baseline extraction method. Evaluators trained in medical informatics are marked with an asterisk.

Specialty	Recall	Precision	F-score
family physician*	0.723	0.124	0.211
cardiologist*	0.447	0.114	0.181
plastic surgeon*	0.827	0.101	0.179
internist*	0.565	0.124	0.204
general surgeon	0.333	0.075	0.122
medical imaging	0.917	0.099	0.179
Average	0.635	0.106	0.182

5. Discussion

The results of this baseline pilot evaluation are encouraging. Similarly to Declerck and Alcantara (2006) who identified the title, caption, and abstract of a Web document among the text regions possibly relevant to image annotation, we found captions, mentions, abstracts and titles of scientific publications to provide sufficient information for image annotation. Although the

information was easily recognized by the evaluators, on average, only 64% of the desirable indexing terms could be found using the existing extraction methods and ontologies. More sophisticated mapping algorithms are needed to extract another 15% of the terms, and more complex natural language processing and ontology expansion are needed to identify the remaining terms.

The pilot evaluation clearly indicates that although there is some correlation between the MeSH terms assigned to a paper and image annotation, only a small proportion of the MeSH terms could be used to describe an image, and additional indexing terms have to be extracted from the text.

The variations in the annotation results among the annotators could be partially attributed to the underspecified image annotation rules. The small number of the images annotated by more than one evaluator does not allow computing inter-annotator agreement scores, but there are indications that the differences could be reduced by better defined rules. For example, in one case, two evaluators marked the extracted term *Hypertrophic Cardiomyopathy* as useful, but only one of them also rated *Echocardiography* as a useful term. Had the instructions clearly stated that if a term belongs to the coarse-level annotation, it should not be used for the medium-level description, the discrepancy might have been avoided. We plan to develop a set of specific rules that describe the appropriate terminology, annotation precision, etc. as described in (Grubinger et al., 2006).

6. Future work

In the next phase, we will focus on the improvement of the evaluation/annotation interface; improvement of the coarse-level controlled vocabularies; selection of the extracted terms to be suggested as indexing terms; improvement of term extraction, and expansion of the test collection. The implementation of some of the improvements to the interface and coarse-level vocabularies suggested by the evaluators is already underway. Figures 3 presents the changes to the coarse-level annotation tab implemented after the pilot evaluation. The changes involve a better layout, a search function for controlled vocabulary terms for coarse level anatomy annotation, and a new teaching quality annotation axis.

7. Acknowledgements

This study was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

8. References

- Sameer K. Antani, Dina Demner-Fushman, Jiang Li, Balaji V. Srinivasan, and George R. Thoma. 2008. Exploring use of images in clinical articles for decision support in Evidence-Based Medicine. *In Proceedings of the 20th SPIE/IS&T Electronic Imaging Conference*, pages 1–10.

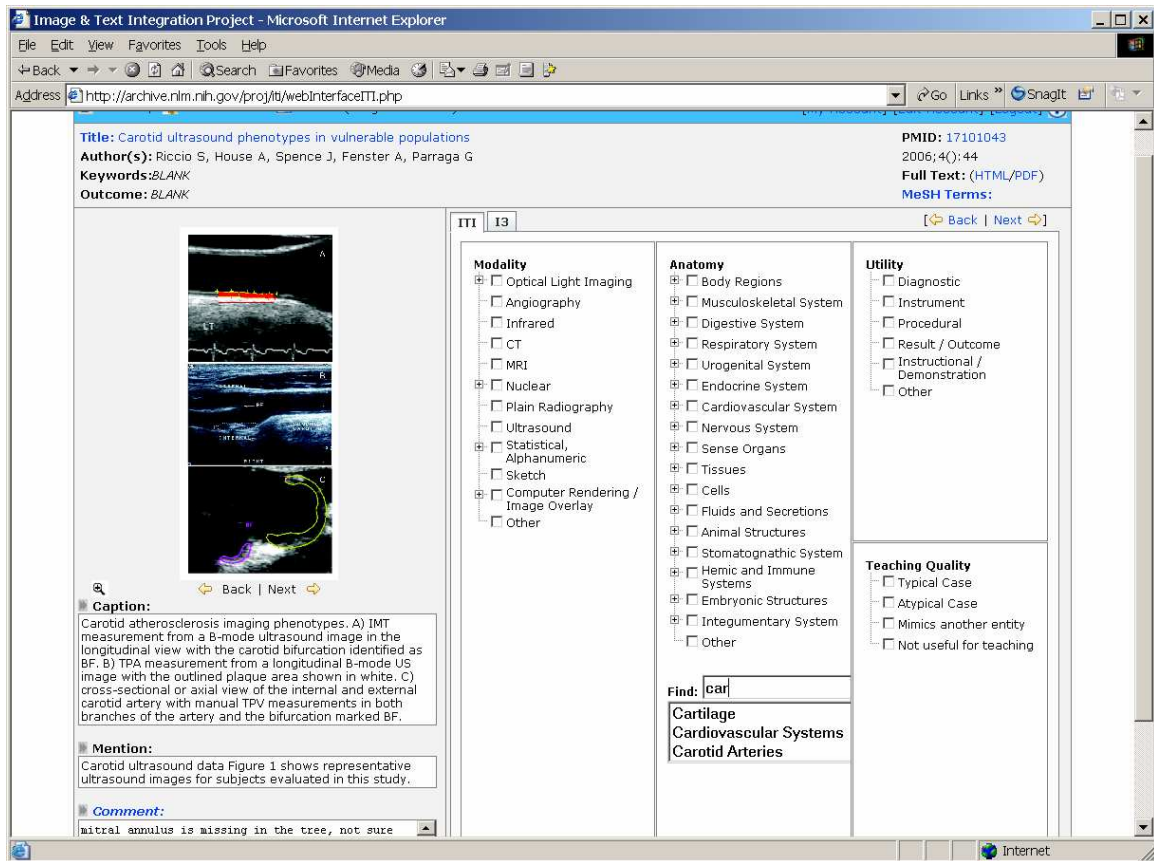


Figure 3: Modified evaluation interface following the evaluators' feedback.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *In Proceedings of the 2001 Annual Symposium of the American Medical Informatics Association (AMIA 2001)*, pages 17–21.

Thierry Declerck and Manuel Alcantara. 2006. Semantic analysis of text regions surrounding images in Web documents. *In OntoImage 2006 Workshop on Language Resources for Content-based Image Retrieval*, pages 9–12.

Dina Demner-Fushman, Sameer K. Antani, and George R. Thoma. 2007. Automatically finding images for clinical decision support. *In Proceedings of the IEEE Workshop on Data Mining in Medicine (DMMed '07)*, pages 139–144.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. *In OntoImage 2006 Workshop on Language Resources for Content-based Image Retrieval*, pages 13–23.

William Hsu, L. Rodney Long, and Sameer K. Antani. 2007. SPIRS: A framework for content-based image retrieval from large biomedical databases. *In Proceedings of the Medinfo Congress*, pages 188–192.

Zhiyun Xue, Sameer K. Antani, L. Rodney Long, Jose Jeronimo, and George R. Thoma. 2007. Investigating CBIR techniques for cervicographic images. *In Proceedings of the 2007 Annual Symposium of the American Medical Information Association (AMIA 2007)*, pages 826–830.

Jian Yao, Sameer K. Antani, L. Rodney Long, George R. Thoma, and Zhongfei Zhang. 2006. Automatic medical image annotation and retrieval using SECC. *In Proceedings of the 19th International Symposium on Computer-Based Medical Systems (CBMS 2006)*, pages 820–825.

Object/Background Scene Joint Classification in Photographs Using Linguistic Statistics from the Web

Bertrand Delezoide, Guillaume Pitel, Hervé Le Borgne

Gregory Greffentette, Pierre-Alain Moëllic, Christophe Millet

CEA/LIST

Centre de Fontenay aux Roses

BP 6 92265 Fontenay-aux-Roses

bertrand.delezoide@cea.fr, guillaume.pitel@cea.fr, herve.le-borgne@cea.fr

gregory.greffentette@cea.fr, pierre-alain.moellic@cea.fr, christophe.millet@cea.fr

Abstract

Object and scene recognition is widely recognized as a difficult problem in computer vision. We present here an approach to this problem that merges recognition of an object and its background. Relying on the assumption that given objects are strongly linked to given background scenes (a deer is more likely to appear in a forest than on an iceberg), we learn object classifiers using joint estimations of object and scene. Such an approach would normally require a large quantity of training images labelled with object/background scene associations. To circumvent costly manual training set labelling, we propose a cross-modal approach, learning and incorporating contextual information via automatic text analysis from the Web, to generate the conditional probabilities of an object given a background scene. This method allows us to strictly distinguish the object classifier from the background scene classifier, and then merge them using estimated conditional probabilities through a learned Bayesian network. The key contribution of this paper is a framework that provides a unified, multimodal approach to learning and using contextual information for improving image processing using statistics obtained from processing Web text.

1. Introduction

Classifying objects and background scenes is a challenging task, in particular because of the ambiguities in the appearance of visual data. As a source of useful information to tackle this issue, one can distinguish *appearance* and *context*. In this paper, appearance information refers to the features commonly used for objects and scene recognition such as color and texture histogram. On the other side the context refers to the information relevant to the detection task but not directly due to the physical appearance of the object, such as their semantic nature or their relative position and scale (Wolf and Bileschi, 2006). In other words, the context can be seen as an expression of the particular relationship that link an object and the background within a natural image. It well worth noting that several evidences coming from neuroscience have shown that human strongly rely on the context to recognize objects (Cox et al., 2004).

The use of contextual information for classification has already been successfully considered using fusion frameworks learned on visual information from annotated images corpora (Luo and Savakis, 2001)(Torralba et al., 2004)(Jasinschi et al., 2002)(Giridharan et al., 2002). This type of joint estimation relies on learning the co-occurrence of a given object with all the possible types of backgrounds within the images. Note that the learning database must contain a significant number of all the possible object/background associations. Such corpora exist for specific domains but are very expensive to build in general. Most of the existing annotated corpora have a unique annotation per image, considering specifically a given object without annotating the background (Fei-Fei et al., 2004)(Everingham et al., 2006) or the contrary. Moreover the usual size of those corpora is relatively small. Indeed, for each couple (background, object), one

must collect and annotate a significant amount of images.

The number of association is at least $\max(|background|, |object|)$ (where $| \cdot |$ denotes the number of element of the set) and at most $|background| \times |object|$. The lower bound of this estimation is very unlikely since it would suppose a situation in which a given object always appears in the same background. If one want to jointly annotate background and objects, one has to consider one of the two following solution: 1 - building a "double annotated" base of image; 2 - finding an innovative method to avoid the explicit building of a (double) annotated database of images. We explore this second option on the following.

The key contribution of this paper is a framework that provides a unified approach to learn and incorporate contextual information obtained from automatic text analysis from the Web for object and background scene classification. Using this scheme, one does not need manual annotations of images anymore to learn the contextual relationships between concepts within images. This textual framework is compared to state-of-the-arts frameworks based on BN and Support Vector Machine (SVM) learned on manually annotated corpora. Our new approach shows significant improvement of classification compared to simple non-contextual classification and gets closer from the performances obtained by the most efficient frameworks learned on image annotation.

The rest of the paper is organized as follows. The next section deals with the related work on object and background scene classification and contextual-based classification. The image corpus used for the evaluation of our framework and our first classification model of objects and background scenes to evaluate the performances on our testbed is presented in section 3. In section 4, we introduce our new approach for extracting context from the Web as well as the integration framework within the classification process. In section 5, we evaluate our approach on a scene/animal joint

classification problem by comparing its performances to the first classification scheme and to state-of-the-arts contextual models. Concluding remarks and prospective are given in section. 6.

2. Related Work

2.1 Object and background scene categorization

The previous works on recognizing isolated objects of various kinds is mainly divided into two approaches. The first approach localizes potential objects, with an automatic segmentation algorithm, prior to trying to recognize the objects: (Barnard et al., 2002) annotates objects after dividing the image into regions with the normalized cuts segmentation algorithm, then features are computed on each region to allow its classification. The second group recognizes objects without any segmentation step. The most common works in this category are the one based on local features such as object recognition with SIFT features developed by Lowe (Lowe, 1999).

A scene is considered here as the picture of a natural environment such as those taken with usual digital cameras. The problem of *scene categorization* consists in recognizing a very typical environment from the whole image. The first works in this vein focused on problems with a low ambiguity on the concepts to identify such as *natural versus artificial* landscapes (Gorkani and Picard, 1994)(Oliva and Torralba, 2001) or *indoor versus outdoor* scenes (Szummer and Picard, 1998), using a combination of low level features (describing colour and texture) with simple classifiers (such as K-nearest neighbours). They achieved about 90% accurate classification on small databases (from 100 to 1300 images). A step further was proposed in (Vailaya et al., 1998) with a hierarchy among possible categories to classify the scenes (indoor/outdoor, city/landscape, etc). They tested their method on 7000 images and obtained 90% accuracy.

The second approach, generally named *bag of features*, rely on the computation of local features around interest points, then making an aggregative feature (such as a histogram) as a signature of the image. A key challenge is to determine a method to obtain as much robustness as possible in the computation of the local features. A reference in this domain is the SIFT (Lowe, 1999). The last approach, initiated in (Oliva and Torralba, 2001), takes advantage of the statistics of natural images to put into relief some intrinsic properties. Contrary to former approaches that measure the quantity of pre-determined features within each image, this method constructs the image features directly from data. An algorithmic principle, usually linked to some perceptual properties of the human visual system (Hervé Le Borgne, 2007), is applied on a collection of natural scenes to obtain a new basis of representation allowing a particular discrimination between scene categories.

2.2 Contextual Fusion Model

The general idea is to take into account some additional *semantic cues* (sometimes named mid- or high-level features) to classify scenes. Although these extra features are themselves determined from the low level features, the fusion process usually leads to an improvement of the

final classification by considering the global *context* of the scene that express the relationship between the constituting elements. Lots of works exist but one can distinguish two main approaches (see (Bosh et al., 2007) for a review).

The first approach consists in identifying some concepts (*grass, sky, or even indoor or city*) within a region of the image, which can be a segmented object. These concepts are further fused in a general framework that captures scene context by discovering intra-frame as well as inter-frame dependency relations between the semantic concepts. E.g.: Markov Random Fields (MRFs) (Geman and Geman, 1984) or Conditional Random Fields (CRFs) (Torralba et al., 2004). Using a discriminative approach for classification rather than spending the efforts in modeling the generation of the observed data is an advantage of CRFs over the traditional MRFs. The disadvantage of these techniques is that they must consider the relations between all the concepts of the ontology which may make computing time prohibitive.

A solution, given by the second approach, is to create a hierarchy to explicitly represent concepts using a basis of other semantic-concepts. In a similar vein, (Luo and Savakis, 2001)(Jasinski et al., 2002)(Giridharan et al., 2002) consider a set of atomic semantic-concepts such as *sky, music, water, speech*, i.e all those which cannot be decomposed or represented straightforwardly in terms of other concepts. They are assumed to be broad enough to cover the semantic query space of interest. Concepts that can be described in terms of other concepts, such as scenes, are then defined as high-level concepts. Hence, estimation of the scene concepts is a multiclass classification problem over the representation of low-level features and atomic semantic-concepts in a semantic space. It is amenable by the modelling of class conditional densities with Bayesian network (Luo and Savakis, 2001)(Jasinski et al., 2002) or more discriminative techniques such as SVMs (Giridharan et al., 2002). Our approach presented in the Section 5 is based on this hierarchical context modelling.

3. First-level classification

This section deals with the classification without fusion, that is to say with the classification of animals on the one side and the background (scene) classification on the other side. However, since we are finally interested into the joint classification, the database is the same for both types of considered images.

We built our database with images coming from the Web found on Google Image¹². We manually selected 30 categories of animals with 50 images for each animal. The images were then segmented into an object (here the animal) and the background. Six types of background were found (see columns of table 2) among these $30 \times 50 = 1500$ images. Images were segmented using the computer assisted segmentation from the SAIST software developed by Hanbury et al. (Hanbury, 2006). It well worth noting this paper does not deal with the problem of automatic segmentation and thus we used a semi-automatic segmentation in order to specifically

¹²<http://images.google.com/>

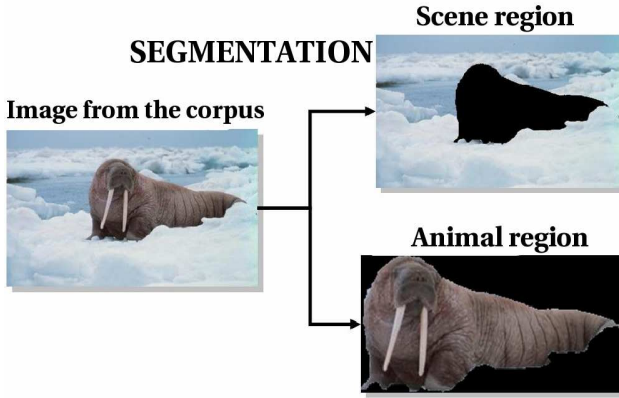


Figure 1: Example of image segmentation from the SAIST software.

study the effect of fusion since it is the topic of the work. In the same vein, the collect of the images was manually checked since we do not study the influence of filtering during this phase. Of course, in a real application these two processing would probably have an influence on the performances. This study is currently in progress and will be reported in further works.

The 1500 images have been randomly separated into a training set of 20 images per animal, and a testing set of 30 images per animals. This random selection has been done 10 times, and the results shown in the following will be an average on these 10 experiments (cross validation). As far as classification is concerned, global features were computed on each region and used to train an SVM classifier. The same method is applied to learn objects and background scenes. Two global features are used: a 64-bins color histogram (RGB quantized into 4 value) and a 512-bins texture histogram (local edge pattern (Cheng and Chen, 2003)). These two features extraction algorithms have been adapted to work on regions with non rectangular shapes, such as the one produced by manual segmentation. It was done considering only pixels within the region for the color histogram, and pixels for which the 8 neighbors are also within the region for the texture histogram.

We combine the color and texture information into 576-bins histograms to learn SVM models with the LibSVM library (Chang and Lin, 2001) with a Gaussian kernel. To manage the multiclass aspect, we used the one-against-one method. The kernels parameters have been estimated by cross-validation on the training data. The result obtained for our baseline is 44.3% of confidence for animals and 50.7% for background scenes.

4. Fusion Models

4.1 General Fusion Scheme

Constructing a generative probabilistic model of image content consists in modeling variables (concept and features) by a general probability distribution able to cover all the possible cases. The distribution then must represent the various descriptions of the image.

Let I be an image of the database; F is the set of features (such as color or texture histograms); A is the semantic concept representing animals presence (e.g. $A = walrus$ or $lion$); S is the concept for

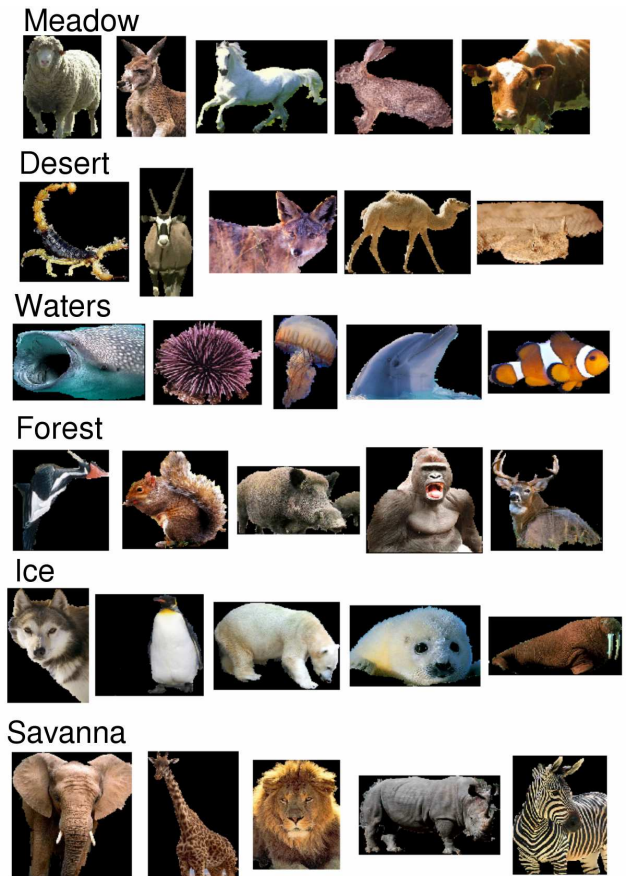


Figure 2: An example of each animal considered in this paper.



Figure 3: Two examples of each background scene considered in this paper.

background scene (e.g. $S = arctic$ or $savanna$). The variables from F are real values, they are said, *observed*, since they are computed by processing of the image without a priori knowledge. The variables from A and S are discrete variables valued in a fixed set (the taxonomy of the concept) and will be evaluated by treatment of the content model. The general classification of the image I consists in attributing the values of the concepts that maximize the probability to observe these concepts knowing the observed variables F . This estimation is the rule of the maximum a posteriori (MAP) noted:

$$\{\hat{S}, \hat{A}\} = \operatorname{argmax}_{S,A} P(S, A | F_S, F_A) \quad (1)$$

Where F_S is the set of features used to classify the scenes and F_A the animals. Then, using the Bayes rule, the MAP rule may be written:

$$\{\hat{S}, \hat{A}\} = \operatorname{argmax}_{S,A} P(S, A, F_S, F_A) \quad (2)$$

The expression of the general joint probability of the random variables is fairly complex. A simplifying method consists in restricting the model structure in order to express the joint probability by several independent terms. The main idea of this method is to specify a number of probabilistic dependences between random variables, based on the a priori knowledge of the modeled phenomenon. That allows reducing the complexity of the inference and learning in comparison with a model where all the probabilistic dependences are considered. In this case, the classification scheme without fusion presented in the fourth section may be approached by considering that the animals and the scene are statistically independent. The MAP rule may then be expressed by the maximization of two independent terms:

$$\{\hat{S}, \hat{A}\} = \operatorname{argmax}_{S,A} P(S | F_S) P(A | F_A) \quad (3)$$

$$\{\hat{S}, \hat{A}\} = \{\operatorname{argmax}_S P_S, \operatorname{argmax}_A P_A\} \quad (4)$$

Where P_S and P_A are the probability of the concepts knowing the associated features calculated by the first SVM classification. Our first assumption is that the independence hypothesis is too strong and that considering the dependence relationships between concepts help to better understand the context of a picture and then improves classification performances.

4.2 SVM Late-Fusion Model

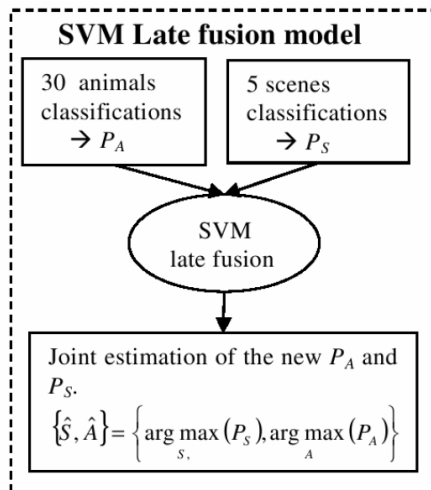


Figure 4: SVM based late fusion model.

The first model is based on SVM Late-Fusion techniques (Westerveld et al., 2003) presented in figure 4. Here, the context of semantic concepts is considered by exploiting concepts interrelation within a pattern recognition problem. Late fusion starts with extraction of low-level features and concepts are learned from these features. Probabilities P_F and P_A are combined afterwards within SVMs models (one for each concept) to yield final

detection probability. Late fusion focuses on the individual strength of concepts within the overall context. A big disadvantage of late fusion schemes is its expensiveness in terms of the learning effort, as the combined representation requires an additional learning stage. Moreover, the second learning phase necessitates an image corpus annotated with all the chosen concepts. For scene extraction this model has shown its efficiency compared to Bayesian Network fusion model. It thus will be considered as a baseline for comparing contextual fusion performances.

4.3 Bayesian Network Fusion Model

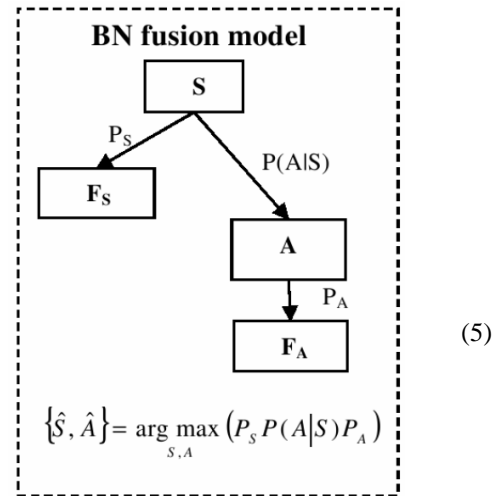


Figure 5: BN based late fusion model.

The second model may be approached by specifying particular probabilistic dependences between the descriptors using a Bayesian Network (BN) formalism. In our case, the BN representing the variables is shown in figure 5. The general joint probability may then be simplified:

$$P(S, A, F_S, F_A) = P(S) P_S P(A | S) P_A \quad (6)$$

If we suppose that the classes from the chosen concepts are equiprobable, the maximum a posteriori may then be expressed by:

$$\{\hat{S}, \hat{A}\} = \operatorname{argmax}_{S,A} P_S P(A | S) P_A \quad (7)$$

The conditional probability of obtaining the animal A knowing the background scene S , $P(A | S)$ is used as a balancing term between the two first probabilities. The evaluation of this conditional probability may be approached by different learning techniques based on external knowledge.

4.3.1 Human Knowledge Technique

A first method consists in manually fixing the conditional probability based on human knowledge. For example, if we assume that a *lion* may not be seen in an *arctic* scene, the probability is arbitrary set to zero: $P(A = \textit{lion} | S = \textit{arctic}) = 0$. During the learning phase, we assume that a particular animal from A_S can only be detected in one particular background scene S :

$P(A_S | \bar{S}) = 0$. We also assume the animals from one scene are equiprobable, that is to say $P(A_S | S) = |A_S| / |S|$ (where $|A_S|$ is the number of animals species that can be found in the background S). This learning technique is easily implementable but rather radical. Assuming that an animal may not be seen in different backgrounds is a deep limit. Moreover, manually fixing the conditional probabilities is feasible in our case where we extract a limited range of concepts but can be problematic when this number grows. This technique can not be considered here, but will serve as a baseline for comparing the others learning techniques.

4.3.2 Annotated Images Corpus Technique

A second technique consists in estimating the conditional probabilities on an annotated images corpus. As concepts are valued in a discrete space, this estimation is based on counting concept values on ground truth images. The joint probability can be computed by counting the frequency of specific configurations among the samples:

$$P(A|S) = \frac{|A \cap S|}{|S|} \quad (8)$$

The main limitation of this approach is that it requires a large images corpus annotated with the whole set of concepts from the chosen ontology. A problem occurs each time one has multiple corpora annotated with a part of the ontology (animals and scenes). It will be necessary to collect a large volume of photographs, with a variety of object/background scene associations. Most of the times, we will have to construct it from scratch by searching the web.

In this article, we propose a third strategy to learn conditional probabilities from external data. This new technique does not need any common images corpora and only uses information automatically extracted from the web. Thus, the learning phase of the fusion scheme does not require manual intervention anymore.

4.4 Joint Probability Estimation from the Web

We have used two resources to count words and cooccurrences: Flickr (Fli,) and Exalead (Exa,).

4.4.1 Web ressources

Flickr is a commercial service for storing and sharing photographs on the internet. One of the main attractive features of this site is the ability to easily tag the photographs. Flickr also proposes two simple search mechanisms: search in tags or in full text descriptions. For each of these modes, usual boolean operators are available: AND, OR, () and NOT.

Exalead is a French search engine that claims to index more than 8 billion pages. Since we use it as a *hit counter*, we preferred it over other popular search engines that may have a bigger coverage, because of the reliability and stability of its count results. Exalead allows for the usual combination of operators to be used in queries: AND, OR, (), NOT, but also more powerful operators such as NEAR (words must be less than 16 words away from each other), NEXT or even OPT for optional words. The NEAR operator is particularly interesting in our case, since we

expected better results from more linguistically-aware counts (co-occurrence in a 16-words window certainly can not be considered as a deep linguistic information, but it is still better than a simple document-based co-occurrence).

4.4.2 Joint Probability Estimation

We queried these engines on 2 parameters: count of individual expressions (object or background-evoking) and count of co-occurring expressions (object/background pairs). We defined four different settings for the query procedure: where co-occurrences are at the document level, ExaleadNear where the pairing queries were made with the NEAR operator, FlickrTags and FlickrText

$$m_0(A, S) = \frac{|A \cap S|}{|A| \cdot |S|} \quad (9)$$

$$P(A|S) = \frac{m_0(A, S)}{\sum_{s \in \text{Scenes}} m_0(A, s)} \quad (10)$$

The quantity we are interested for our purpose is the conditional probability of finding a particular animal given the background scene. While the conditional probability may be a good predictor in the general case, the standard estimation (see equation (8)) is strongly biased toward most frequent animals (in our setting, for instance, horses are cited and photographed more frequently than any other animal). It is highly desirable that the measure be independent of the relative frequency of the animals. For this reason, a measure $m_0(A, S)$ (see equation (9)) close to the Pointwise Mutual Information was used to approximate the conditional probability. We can then define $P(A|S)$ from m_0 with a normalization step (equation (10)).

4.4.3 Example

Class	Terms
Scenes	{ meadow ``green grass" ``tall grass" trunk log branch leaf snow mud tree}, { desert dune oasis sand}, { waters spume plunge dive swim sand}, { forest foliage woods trunk log branch leaf snow mud tree}, { ice floe icefield}, { savanna ``tall grass" ``yellow grass" trunk log branch leaf sand mud tree}
Objects	elephant, horned viper, clownfish, cow, deer, dolphin, dromedary, giraffe, gorilla, hare, horse, husky, jackal, jellyfish, kangaroo, lion, oryx, penguin, polar bear, rhino, boar, scorpion, seal, urchin, sheep, squirrel, walrus, whale, woodpecker, zebra

Table 1: Terms or group of terms used for joint probability estimation.

In our experiment, we approximate the relation between animals and scenes. We count individual and (animal/scene) joint count on the terms presented in table 1. Based on the counts we realized using the Exalead search engine and the NEAR operator to join animal and scene terms, the joint probabilities we obtain are

presented in table 2. As shown by the urchin example, some estimations can be totally wrong, perhaps because of occasional odd answers from the search engine. This is however definitely cheaper and faster to collect a set of terms describing scenes and animals than to collect a collection of pictures representing the “natural” distribution of animals in different environments.

	meadow	desert	waters	forest	ice	savanna
elephant	.22	.18	.05	.16	.06	.33
horned viper	.04	.62	.17	.03	.00	.13
clownfish	.04	.11	.72	.05	.02	.07
cow	.41	.09	.08	.12	.11	.19
deer	.20	.08	.03	.42	.12	.15
dolphin	.06	.14	.60	.07	.08	.05
dromedary	.04	.70	.09	.04	.05	.07
giraffe	.17	.05	.03	.10	.03	.62
gorilla	.07	.57	.08	.08	.11	.09
hare	.29	.14	.10	.13	.07	.26
horse	.16	.09	.10	.18	.36	.12
husky	.18	.04	.04	.15	.48	.12
jackal	.18	.37	.06	.10	.06	.23
jellyfish	.05	.14	.49	.05	.19	.08
kangaroo	.16	.19	.10	.35	.07	.13
lion	.24	.14	.09	.10	.13	.31
oryx	.06	.44	.06	.03	.01	.41
penguin	.09	.05	.13	.09	.56	.08
polar bear	.01	.00	.07	.01	.90	.00
rhino	.24	.17	.08	.11	.12	.27
boar	.23	.14	.08	.24	.08	.23
scorpion	.10	.33	.12	.11	.11	.22
seal	.08	.07	.18	.08	.49	.09
urchin	.03	.10	.16	.04	.01	.66
sheep	.30	.16	.05	.10	.25	.15
squirrel	.24	.08	.08	.26	.14	.20
walrus	.01	.05	.02	.01	.91	.01
whale	.04	.09	.52	.05	.25	.05
woodpecker	.26	.10	.05	.29	.08	.21
zebra	.26	.09	.08	.08	.10	.39

Table 2: Animal/Scene joint probability estimation using Exalead and NEAR operator.

5. Experiments

The confidences in the classifications of animals and their associated scenes are presented in table 2. We compare the classification performances of the different fusion models: classification without fusion (No fusion), BN fusion learned on images corpus (BNima), on Exalead cooccurrences (BNexa), on ExaleadNear (BNexan), on FlickrTags (BNftag), on Flickrtexts (BNftxt) and SVM late fusion (SVMlate).

This experiment gives rise to two interesting results. First, contextual fusion can be used to improve classification performances. Second, conditional probability learned from the WWW provides useful information for the joint estimation of animals and scenes.

SVM late fusion models better consider the correlation between the classification scores. This is due to the quality of the estimation of their inter-relation, learned from the ground truth examples from photographs mapped in the initial semantic space through the kernel function. It thus reaches the best classification

performances. The results on BN demonstrates their ability to handle context in the images and shows the best performances of BN fusion model by learning the context from ground truth. BN learned on the Web is less efficient, but still shows a fair improvement compared to the classification scheme without fusion (+5.3% on average for BNftxt). These results demonstrate that contextual fusion using information extracted from the Web is efficient. The main advantage of our method is to circumvent costly manual training set labelling of images. This method allows us to strictly distinguish the object classifier from the background scene classifier, and then merge them using estimated conditional probabilities through an easily learned Bayesian network via automatic text analysis from the Web.

Within the different BN fusion models learned from the Web, we observe a variation of classification performances. The performances are always lower than the one of the BN model learned on the image corpus. Indeed, it seems that the more a BN fusion is efficient, the more conditional probabilities are close from the image ground truth.

Our next goal will then be to enforce the robustness of joint probability estimation from the Web in order to get closer from the estimation obtained with image corpora. Another way of improvement would be to better considerate the statistical dependence relationships between animals and scenes. Indeed BN model is less efficient than SVM for this task, as BN only consider a first order relationship through the conditionals probabilities of observing animals knowing the scenes. Another fusion framework should be found to obtain both *good statistical dependence considering* and *Web-based learning phase*.

	No fusion	BNima	BNexa	BNexan
Animals	44.3	49.1	45.5	47.5
Scenes	50.7	64.2	54.1	57.8

	BNftag	BNftxt	SVM Late
Animals	46.1	47.6	52.9
Scenes	54.5	58.0	67.6

Table 3: Classification performances of the fusion models

6. Conclusion

In this article, we have addressed the problem of objects and background scenes joint classification from consumer photograph using contextual information. We proposed to learn a Bayesian Network fusion model with information extracted from the Web, instead of annotated images. This new model leads to drastically reduce the manual annotation effort that is a critical task to test classification fusion models. Feasibility of such a framework was demonstrated for the automatic annotation of photographs with animals and background scenes concepts.

A fair improvement compared to the classification results obtained without fusion (+5% precision) shown the efficiency of our method. Using our method, one can now consider to efficiently learn fusion schemes to automatically annotate photographs using large

ontologies (such as LSCOM), or very specialized ones. Several directions exist to improve the classifications fusion scheme described in this article. First, joint probability extraction from the Web may be developed to get closer from ground truth from the images. Secondly, an alternative fusion framework could be considered in order to better model the dependencies between objects and scenes within joint classification scheme.

7. Acknowledgments

This work is sponsored by the European Network of Excellence MUSCLE¹³. We also thank the Direction Generale des Entreprises for funding us through the regional business cluster Systematic (project POPS¹⁴).

8. References

- Kobus Barnard and Pinar Duygulu and Nando de Freitas and David Forsyth and David Blei and Michael I Jordan. Matching Words and Pictures. *Journal of Machine Learning Research, Special Issue on Text and Images*, 3:1107--1135, 2002.
- Bosh, Anna and Munoz, Xavier and Marti, Robert. Which is the best way to organize/classify images by content?. *Image and Vision Computing*, 25(6):778--791, 2007.
- Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a Library for Support Vector Machines*. 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
<http://www.flickr.com>.
<http://www.exalead.com>.
- Cheng, Ya-Chun and Chen, Shu-Yuan. Image classification using color, texture and regions. *Image Vision Computing*, 21(9):759--776, 2003.
- Cox, David and Meyers, Ethan and Sinha, Pawan. Contextually evoked object specific responses in human visual cortex. *Science*, 304:115--117, 2004.
- M. Everingham and A. Zisserman and C. Williams and L. Van Gool and M. Allan and C. Bishop and O. Chapelle and N. Dalal and T. Deselaers and G. Dorko and S. Duffner and J. Eichhorn and J. Farquhar and M. Fritz and C. Garcia and T. Griffiths and F. Jurie and D. Keysers and M. Koskela and J. Laaksonen and D. Larlus and B. Leibe and H. Meng and H. Ney and B. Schiele and C. Schmid and E. Seemann and J. Shawe-Taylor and A. Storkey and S. Szedmak and B. Triggs and I. Ulusoy and V. Viitaniemi and J. Zhang. The 2005 PASCAL Visual Object Classes Challenge. *Selected Proceedings of the First PASCAL Challenges Workshop, LNAI, Springer-Verlag*, 2006.
- L. Fei-Fei and R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *Comp. Vis. and Pattern Recogn. (CVPR) 2004, Workshop on Generative-Model Based Vision*, 2004.
- Geman, Stuart and Geman, Donald. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721--741, 1984.
- Iyengar Giridharan and Nock, Harriet J. and Neti Chalapathy and Franz, Martin. Semantic indexing of multimedia using audio, text and visual cues. *Proceedings of IEEE International Conference on Multimedia and Expo 2002 (ICME '02)*, 2002.
- Gorkani, Monika M. and Picard, Rosalind W. Texture Orientation For Sorting Photos "At A Glance". *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, pages 459-464, 1994.
- Allan Hanbury. Review of image annotation for the evaluation of computer vision algorithms. Technical report, PRIP-TR-102, PRIP, T.U. Wien., 2006.
- Hervé Le Borgne, Anne Guérin-Dugué, Noel E. O'Connor. Learning Mid-level Image Features for Natural Scene and Texture Classification. *IEEE transaction on Circuits and Systems for Video Technology*, 17(3):286--297, 2007.
- Jasinschi, Radu and Dimitrova, Nevenka and McGee, Thomas and Agnihotri, Lalitha and Zimmerman, John and Li, Dongge and Louie, Jennifer. A probabilistic layered framework for integrating multimedia content and context information. *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, 2002.
- David G. Lowe. Object Recognition from Local Scale-Invariant Features. *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1150-1157, 1999.
- Luo, Jiebo and Savakis, Andreas E. Indoor vs. Outdoor Classification of Consumer Photographs Using Low-level and Semantic Features. *Proceedings of International Conference on Image Processing' 01*, pages 745--748, 2001.
- Aude Oliva and Antonio B. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145-175, 2001.
- Martin Szummer and Rosalind W. Picard. Indoor-Outdoor Image Classification. *IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*, pages 42-51, 1998.
- Torralba, Antonio and Murphy, Kevin P. and Freeman, William T. Contextual models for object detection using boosted random fields. *Proceedings of Advances in Neural Information Processing Systems 17 (NIPS 2004)*, 2004.
- Aditya Vailaya and Anil Jain and Hong Jiang Zhang. On image classification: city images vs. landscapes. *Pattern Recognition*, 31(12):1921--1935, 1998.
- Thijs Westerveld and Arjen P. de Vries and Alex van Ballegooij and Franciska de Jong and Djoerd Hiemstra. A probabilistic multimedia retrieval model and its evaluation. *EURASIP Journal on Applied Signal Processing, Special issue on Unstructured Information Management from Multimedia Data Sources*, 2003(2):186--198, 2003.
- Wolf, Lior and Bileschi, Stanley. A Critical View of Context. *International Journal of Computer Vision*, 69(2):251--261, 2006.

¹³<http://www.muscle-noe.org/>

¹⁴<http://www.pops-systematic.org/>

Identifying News Broadcasters' Ideological Perspectives Using a Large-Scale Video Ontology

Wei-Hao Lin and Alexander Hauptmann

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.
{whlin,alex+}@cs.cmu.edu

Abstract

Television news has been the predominant way of understanding the world around us, but individual news broadcasters can frame or mislead audience's understanding about political and social issues. We aim to develop a computer system that can automatically identify highly biased television news, which may prompt audience to seek news stories from contrasting viewpoints. But can computers determine if news videos were produced by broadcasters holding differing ideological beliefs? We developed a method of identifying differing ideological perspectives based on a large-scale visual concept ontology, and the experimental results were promising.

1. Introduction

Television news has been the predominant way of understanding the world around us. Individual news broadcasters, however, can frame, even mislead, audience's understanding about political and social issues. A recent study shows that people's main news sources are highly correlated with their misconceptions about the Iraq War (Kull, 2003). 80% of the respondents whose primary news source is FOX have one or more misconceptions, while among people whose primary source is CNN, 50% have misconceptions.

The difference in framing news events is clearer when we compare news broadcasters across national and language boundaries. For example, Figure 1 shows how an American broadcaster (NBC) and an Arabic broadcaster (LBC) portray Yasser Arafat's death in 2004. The two broadcasters' footage looks very different: NBC shows stock footage of Arafat, while LBC shows the actual funeral and interviews with general public.

We consider a broadcaster's bias in portraying a news event "ideological." We take the definition of ideology as "a set of general beliefs socially shared by a group of people" (van Dijk, 1998). Television news production involves a large number of people who share similar social and professional beliefs. A news broadcaster may consistently exhibit bias in reporting political and social issues partly because producers, editors, and reporters collectively make similar decisions (e.g., what to cover, who to interview, and what to show on a screen) based on shared value judgments and beliefs.

We aim to develop a computer system that can automatically identify highly biased television news. Such system may increase audience's awareness about individual news broadcasters' bias and prompt them to seek news stories from contrasting viewpoints. However, can computer automatically understand differing ideological perspectives expressed in television news footage?

- In this paper we proposed a method of identifying differing ideological perspectives in news video based on the imagery chosen to show on the screen. We motivated our method based on visual concepts in Section 2.. We described how to represent a video in terms of visual concepts (e.g., outdoor, car, and people walking) in Section 3.1., and then how to quantify the similarity between two news video footage in terms of visual concepts in Section 3.2..
- We evaluated the proposed method on a large broadcast news video archive (Section 4.1.). To determine if two videos portray the same news event from differing ideological perspectives, we trained a classifier to make a binary decision (i.e., same perspective or different perspectives). The classifier was shown to achieve high accuracy in Section 4.3.. We applied the same idea to determine if two videos covered the same news event in Section 4.2..
- So far we conducted the experiments using manual concept annotation to avoid concept classifiers' poor performance being a confounding factor. In Section 4.4. we repeated the above experiments and replaced manual annotations with empirically trained concept classifiers.

2. Motivation

We were inspired by the recent work on developing large-scale concept ontology for video retrieval (Hauptmann, 2004), and considered a specific kind of visual grammar that may exhibit ideological perspective: composition (Efron, 1972). Here visual concepts are generic objects, scenes, and activities (e.g., outdoor, car, and people walking). Visual concepts can represent a video's visual content more closely than low-level features (e.g., color, texture, and shape) can. Many researchers have actively developed concept classifiers to automatically detect concepts' presence in video. A concept classifier reads an image and outputs the likelihood that a visual concept is present on the screen. Therefore, if computers can automatically identify the visual concepts, computers may be able to learn the



(a) From an American news broadcaster, NBC



(b) From an Arabic news broadcaster, LBC

Figure 1: The key frames of the television news footage about Yasser Arafat’s death from two broadcasters.

difference between broadcasters holding differing ideological perspectives based on what are chosen to show in news footage.



(a) CNN



(b) LBC

Figure 2: The text clouds showed the frequency of the visual concepts that were chosen by two broadcasters in the Iraq War stories. The larger a visual concept, the more frequently the concept was shown in news footage.

We illustrate the idea in Figure 2. We counted the visual concepts in the television news footage about the Iraq War from two different broadcasters (an American broadcaster CNN vs. an Arabic broadcaster LBC), and displayed them in text clouds (see Section 4.1. for more details about the data). Due to the nature of broadcast news, it is not surprising to see many people-related visual concepts (e.g., “Adult”, “Face”, and “Person”). Because the news stories are about the Iraq War, it is also not surprising to see many war-related concepts (e.g., “Weapons”, “Military Personnel”, and “Daytime Outdoor”). The surprising differences, however, lie in the subtle emphasis on some concepts. “Weapons” and “Machine Guns” are shown more often in CNN (relative to other visual concepts in CNN) than in LBC. On the contrary, “Civilian Person” and “Crowd” are shown more often in LBC than in CNN. How frequently some visual concepts are chosen seems to reflect a broadcaster’s ideological perspective on a particular news event.

3. Measuring Semantic Similarity in Visual Content

To develop a computer program that can identify videos conveying differing ideological perspectives on a news event, we need to address the following two questions:

1. Can computers determine if two television news stories are about the same news event?
2. Given two television news stories on the same

news event, can computers determine if they portray the event from differing ideological perspectives?

We could identify news stories’ topic using textual clues (e.g., words in automatic speech recognition transcripts), but here we attack a more challenging question: grouping television news stories on the same event using only visual clues. More and more videos are produced and consumed by users on the Internet. Contrary to news videos, web videos do not usually come with clear voice-over that describes what a video is about. An imagery-based topic tracking approach is more likely to be applicable for web videos than a text-based approach. The two research questions can be boiled down to the same question:

How well can we measure the similarity in visual content between two television news videos?

News videos on the same news event are likely to have similar visual content, while news videos on different news events are less likely to have similar visual content. Similarly, given two news videos on the same news event, broadcasters holding similar ideological beliefs are likely to portray the new event in a similar manner, while news broadcasters holding different ideological views are less likely to display similar visual content. Therefore, the key research question becomes measuring the “semantic” similarity in visual content.

3.1 Representing Video As Visual Concepts

We proposed a method of measuring semantic similarity between two news stories using a large-scale visual concept ontology. Our method consists of four steps, as illustrated in Figure 3. In Step 1 we first run a shot detector to detect shot boundaries in a news story, and select the middle frame of a shot as its key frame. In Step 2 we check if any concepts in a visual concept ontology are present in the key frames. A concept’s presence can be manually labeled by human annotators, but can be also automatically but less accurately labeled using machine learning classifiers. An example key frame and its visual concepts are shown in Figure 4.

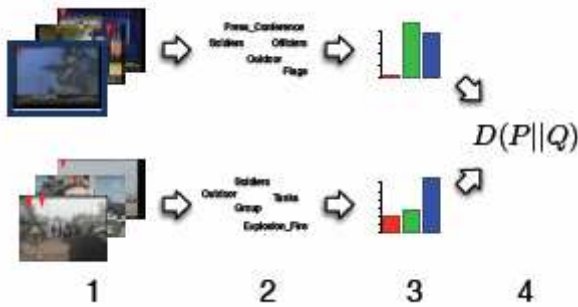


Figure 3: Our method of measuring similarity in visual content consisted of four steps. Step 1: extract videos’ key frames. Step 2: determine what visual concepts are present in key frames. Step 3: model the occurrences of visual concepts using a multinomial distribution. Step 4: measure “distance” between two multinomial distributions using Kullback-Leibler divergence.



Figure 4: This key frame is annotated with the following LSCOM visual concepts: Vehicle, Armed Person, Sky, Outdoor, Desert, Armored Vehicles, Daytime Outdoor, Machine Guns, Tanks, Weapons, Ground Vehicles.

We choose to represent the visual content of a television news story as a set of visual concepts shown on the screen. By visual concepts we mean generic objects, scenes, and activities (e.g., outdoor, car, and people walking). Low-level features (e.g., color, texture, shape) are easy to compute but fail to closely represent a video’s visual content. For example, to compare how different broadcasters portray the Iraq War, knowing how many “soldiers” (a visual concept) they choose to show is much more informative than knowing how many brown patches (a low-level color feature) are shown.

Category	Examples
Program	advertisement, baseball, weather news
Scene	indoors, outdoors, road, mountain
People	NBA players, officer, Pope
Objects	rabbit, car, airplane, bus, boat
Activities	walking, women dancing, cheering
Events	crash, explosion, gun shot
Graphics	weather map, NBA scores, schedule

Table 1: The major categories and sample LSCOM concepts in each category.

In this paper we chose the Large-Scale Concept Ontology for Multimedia (LSCOM) (Kennedy and Hauptmann, 2006) to represent television video’s visual content. LSCOM, initially developed for improving video retrieval, contains hundreds of generic activities, objects, and scenes¹⁵. LSCOM started from more than ten

¹⁵ The complete list of visual concepts is available at <http://www.lscm.org/concept.htm>

thousands of concepts collected from various sources such as TGM, Time Life, TV Anytime, Comstock, and WordNet. Later around one thousand concepts were chosen based on video retrieval utility, machine-learning feasibility, and observability. The LSCOM taxonomy was also mapped to Cyc to suggest new concepts. The major categories and example concepts in each category are listed in Table 3.1..

3.2 Measuring Similarity using Visual Concept Representation

In Step 3 we model the occurrences of visual concepts in key frames using a statistical distribution. A natural choice for discrete occurrences is a multinomial distribution. We take the visual concepts detected in Step 2, and count how many times every concept in a visual concept ontology appears. We obtain the maximum likelihood estimate (MLE) of a multinomial distribution’s parameter by dividing the visual concept frequency by the total number of visual concepts in a news video. Because the number of unique visual concepts in a news story is usually much smaller than the total number of concepts of a visual concept ontology, the MLE contains many zero entries. We thus smooth the MLE by adding a small pseudo count (0.001), which is equal to the maximum posteriori estimate with a Beta prior (Manning and Schütze, 1999). We measure the similarity between two videos’ multinomial distributions in terms of Kullback-Leibler (KL) divergence (Cover and Thomas, 1991). KL divergence is commonly used to measure the “distance” between two statistical distributions. The KL divergence between two multinomial distributions P and Q is defined as follows:

$$D(P||Q) = \sum_c P(c) \log \frac{P(c)}{Q(c)},$$

where c is all visual concepts. The value of KL divergence quantifies the similarity between two news videos in terms of visual concepts chosen by individual broadcasters. The smaller the value of KL divergence, the more similar two news videos. KL divergence is asymmetric, and we take the average of $D(P ||Q)$ and $D(Q||P)$ as the (symmetric) distance between P and Q.

4. Measuring Semantic Similarity in Visual Content

4.1 Data

We evaluated the proposed method of identifying differing ideological perspectives on a broadcast news video archive from the 2005 TREC Video Evaluation (TRECVID) (Over et al., 2005). The TRECVID 2005 video archive consisted of television news videos recorded in late 2004. The news programs came from multiple news broadcasters in three languages: Arabic, Chinese, and English, as shown in Table 2.

Language	Hours	News Broadcasters
Arabic	33	LBC
Chinese	52	CCTV, NTDTV
English	73	CNN, NBC, MSNBC

Table 1: The news broadcasters and the total length of newsvideos in each language in the TRECVID’05 video archive.

We used the official shot boundaries that the TRECVID organizer, NIST, provided for the TRECVID 2005 participants. We ran an in-house story segmentation program to detect news story boundaries (Hauptmann et al., 2005), resulting in 4436 news stories. The story segmentation program detected a news story’s boundary using cues such as an anchor’s presence, commercials, color coherence, and average story length. We removed anchor and commercial shots because they contained mostly talking heads and conveyed little ideological perspective. We collected ten news events in late 2004 and news videos covering these news events. We made sure the news events in Table 3 were covered by broadcasters in more than one language. A news story covered a news event if a news story’s keywords were mentioned in the video’s English automatic speech recognition (ASR) transcripts. NIST provided English translation for non-English news programs. Note that ASR transcripts were used only for linking stories on the same news event. LSCOM annotators did not use ASR transcripts and made judgments solely based on visual content.

News Event	Hours
Iraq War	231
United States presidential election	114
Arafat’s health	308
Ukrainian presidential election	11
AIDS	21
Afghanistan situation	42
Tel Aviv suicide bomb	2
Powell’s resignation	45
Iranian nuclear weapon	46
North Korea nuclear issue	51

Table 3: The number of television news stories on the ten news events in late 2004.

We used visual concepts annotation from the Large-Scale Concept Ontology for Multimedia (LSCOM) v1.0 (Kennedy and Hauptmann, 2006). The LSCOM annotations consisted of the presence of each of the 449 LSCOM visual concepts in every video shot of the TRECVID 2005 videos. There are a total of 689064 annotations for the 61901 shots, and the median number of annotations per shot is 10.

We conducted the experiments first using the LSCOM annotations, and later replaced manual annotations with predictions from empirically trained concept classifiers. Using manual annotations is equal to using very accurate concept classifiers. Given the state-of-the-art classifiers for

most visual concepts are far from perfect, why would we start from assuming perfect concept classifiers? It is because manual annotations allow us to test the idea of measuring similarity in visual concept using concepts without being confounded by the poor accuracy of the concept classifiers.

4.1 Identifying News Videos on the Same News Event

Because we are interested in how the same news event is portrayed by different broadcasters, we need to find the television news stories on the same news event in a video archive. As we argued in Section 3., this task boils down to comparing similarity between two videos’ visual content. News videos on the same news event are likely to show similar visual content. Given two news videos, we could measure their similarity in terms of visual concepts as proposed in Section 3..

We developed a classification task to evaluate the proposed method of identifying news videos on the same event. Each time the classifier is presented with a pair of television news videos, and is asked to make a binary decision between two categories: Different News Events (DNE) vs. Same News Event (SNE). DNE contains news video pairs that are from the same broadcaster but on different news events (e.g., two videos from CNN: one is about the “Iraq War” and the other is about “Powell’s resignation”). SNE contains news video pairs from the same broadcaster and on the same news event (e.g., two videos from CCTV about the same event “Tel Aviv bomb”). The predictor for the classification task is the value of KL divergence between two videos. Our method is effective if such classifier achieves high accuracy.

Among all possible video pairs that satisfy the conditions of Different News Event (DNE) and Same News Event (SNE), we randomly sampled 1000 video pairs for each category. We looked up their LSCOM concept annotations (Section 3.1.), estimated multinomial distributions’ parameters, and trained classifiers based on the values of (symmetric) KL divergence (see Section 3.2.). We varied the training data from 10% to 90%, and reported the accuracy on the held-out 10% of video pairs. Accuracy is defined as the number of video pairs that are correctly classified divided by the total number of video pairs in the held-out set. Because there were an equivalent number of video pairs in each category, a random guessing baseline would have 50% accuracy. We repeated the experiments 100 times by sampling different video pairs, and reported the average accuracy. The choice of classifier did not change the results much, and we reported only the results using Linear Discriminant Analysis and omitted the results using Support Vector Machines.

The experimental results in Figure 5 showed that our method based on visual concepts can effectively tell newsvideos on the same news event from news videos on different news events. The classification accuracy was significantly better than the random baseline (t-test, $p < 0:01$), and reached a plateau around 70%. Our concept-based method of identifying television news stories on the same event could thus well complement other methods based on text (Allan, 2002; Zhang et al., 2004), color (Zhai and Shah, 2005), and near-duplicates

images (Wu et al., 2007). Although LSCOM was initially developed for supporting video retrieval, the results also suggested that LSCOM contained large and rich enough concepts to differentiate news videos on a variety of news events.

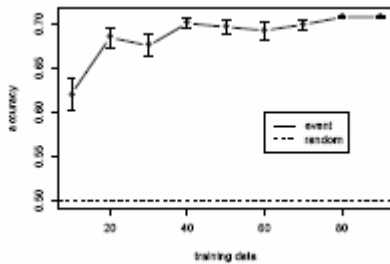


Figure 5: The proposed method can differentiate news video pairs on the same news events from the news video pairs on different news events significantly better than a random baseline. The x axis is the percentage of training data, and the y axis is the binary classification accuracy.

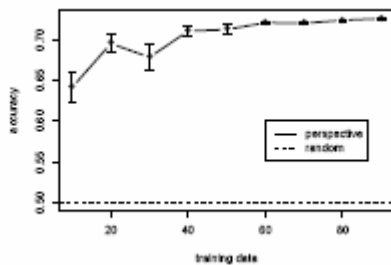


Figure 6: The proposed method can differentiate the news video pairs conveying the differing ideological perspectives from the news videos conveying similar ideological perspectives significantly better than a random baseline. The x axis is the percentage of training data, and the y axis is the binary classification accuracy.

4.2 Identifying New Videos of Different Ideological Perspectives

Given two news videos on the same news event, how can computers tell if they portray the event from different ideological perspectives? As we hypothesized in Section 2., given a news event, broadcasters holding similar ideological beliefs (i.e., the same broadcaster) are likely to choose similar visual concepts to compose news footage, while broadcasters holding different ideological beliefs (i.e., different broadcasters) are likely to choose different visual concepts. The task of identifying if two news videos convey differing ideological perspectives boils down to measuring if two videos are similar in terms of visual concepts (Section 3.).

We developed a classification task to evaluate the proposed method of identifying news videos from differing ideological perspectives. There were two categories in the classification task: Different Ideological Perspectives (DIP) vs. Same Ideological Perspectives (SIP). DIP contains news video pairs that are about the same news event and from different broadcasters (e.g., two videos about “Arafat’s death”: one from LBC and one from NBC). SIP contains news video pairs that are about the same event but from the same broadcaster (e.g., two videos both from NTDTV and about “Powell’s resignation”). We trained a binary classifier to predict if a news video pairs belong to DIP or SIP. We followed the

classification training and testing procedure in Section 4.2..

The experimental results in Figure 6 showed that our method based on visual concepts can effectively tell news videos produced by broadcasters holding similar ideological beliefs from those holding differing ideological beliefs. The classification accuracy was significantly better than the random baseline (t-test, $p < 0:01$), and reached a plateau around 72%. Given two news videos are on the same news event, we can then use the propose method to test if they portray the news from differing ideological perspectives.

Because we already knew a video’s broadcaster when the video was recorded, wasn’t the task of identifying if two news videos portray the news event from differing ideological perspectives as trivial as checking if they come from different broadcasters? Although we can accomplish the same task using metadata such as a news video’s broadcaster, this method is unlikely to be applicable to videos that contain little metadata (e.g., web videos on YouTube). We opted for a method of broader generalization, and developed our method solely on visual content and generic visual concepts.

4.3 Concept Classifier’s Accuracy

So far our experiments were based on manual annotations of visual concepts from LSCOM. Using manual annotation is equal to assuming that perfect concept classifiers are available, which is unrealistic given that the state-of-the-art classifiers are far from perfect for most visual concepts (Naphade and Smith, 2004). So how well can computers determine if two news videos convey a differently ideological perspective on a news event using empirically trained classifiers? We obtained 449 LSCOM concept classifiers’ empirical accuracy by training Support Vector Machines on 90% of positive examples and testing on the held-out 10%. We first trained uni-modal concept classifiers using single low-level features (e.g., color histogram in various grid sizes and color spaces, texture, text, audio, etc), and built multimodal classifiers that fused the outputs from best uni-modal classifiers (see (Hauptmann et al., 2005) for more details about the training procedure). We evaluated the performance of the best multi-modal classifiers on the held-out set in terms of average precisions (AP).

We varied concept classifiers’ accuracy by injecting noise into manual annotations. AP is a rank-based evaluation metric, but our experiments relied on set-based metrics. We thus approximated AP using recall-precision break-even points, which was highly correlated with AP (Manning et al., 2008). We randomly flipped the positive and negative labels of visual concepts until we reached the desired breakeven points. We varied the classifiers’ break-even points from APs obtained from empirically trained classifiers to 1.0 (i.e., perfect accuracy), and repeated the experiments in Section 4.2. and Section 4.2.. The experimental results showed that the empirically trained classifiers cannot satisfactorily identify news videos covering the same news event (Figure 7a) and news videos conveying differing perspectives (Figure 7b). Although the classification accuracy using empirically trained concept classifiers (i.e., the leftmost data point) was statistically significantly from random (t-test, $p < 0:01$), the difference was not practically significant. The

median AP of the empirically trained classifiers was 0:0113 (i.e., the x coordinate of the leftmost data point in Figure 7). It was not surprising to see the classification accuracy improved as concept classifiers' break-even points increased. To achieve reasonable performance we seemed to need concept classifiers of break-even points 0:6.

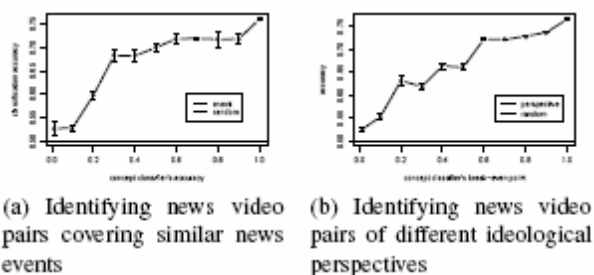


Figure 7: We varied the classifiers' accuracy and repeated the two experiments in Figure 5 and Figure 6. The x axis is the (simulated) classifiers' accuracy in terms of precision-recall break-even points. The leftmost data point was based on the performance of the empirically trained classifiers. The y axis is the classification accuracy.

We should not be easily discouraged by current classifiers' poor performance. With the advance of computation power and statistical learning algorithms, it is likely that concept classifiers' accuracy will be continuously improved. Moreover, we may be able to compensate for poor accuracy by enlarging the number of concepts, as demonstrated recently in the study of improving video retrieval using more than three thousands of visual concepts (Hauptmann et al., 2007).

4. Conclusions

We proposed a method of measuring difference in visual content using a large-scale video concept ontology. The experiment results showed that by representing news footage in terms of visual concepts, we could start to learn news broadcasters' patterns in composing news videos about different news topics and in portraying a news event from different ideological perspectives.

6. Acknowledgements

We would like to thank the anonymous reviewers for their valuable suggestions for improving this paper. This research was supported in part by the National Science Foundation (NSF) under Grant No. IIS-0205219.

7. References

James Allan, editor. 2002. Topic Detection and Tracking: Event-based Information Organization. *Kluwer Academic Publishers*.

Thomas M. Cover and Joy A. Thomas. 1991. Elements of Information Theory. *Wiley-Interscience*.

Edith Efron. 1972. The News Twisters. *Manor Books*.

A. G. Hauptmann, R. Baron, M. Christel, R. Conescu, J. gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang. 2005. CMU Informedia's TRECVID 2005 skirmishes. In *Proceedings of the 2005 TREC Video Retrieval Evaluation*.

Alexander Hauptmann, Rong Yan, and Wei-Hao Lin. 2007. How many high-level concepts will fill the semantic gap in news video retrieval? In *Proceedings of the Sixth International Conference on Image and Video Retrieval (CIVR)*.

Alexander G. Hauptmann. 2004. Towards a large scale concept ontology for broadcast video. In *Proceedings of the Third International Conference on Image and Video Retrieval (CIVR)*.

Lyndon Kennedy and Alexander Hauptmann. 2006. LSCOM lexicon definitions and annotations (version 1.0). Technical Report ADVENT 217-2006-3, Columbia University, March.

Steven Kull. 2003. Misperceptions, the media and the iraq war. http://65.109.167.118/pipa/pdf/oct03/IraqMedia_Oct03_rpt.pdf, October.

Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. *The MIT Press*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. *Cambridge University Press*.

Milind R. Naphade and John R. Smith. 2004. On the detection of semantic concepts at TRECVID. In *Proceedings of the Twelfth ACM International Conference on Multimedia*.

Paul Over, Tzveta Ianeva, Wessel Kraaij, and Alan F. Smeaton. 2005. TRECVID 2005 - an overview. In *Proceedings of the 2005 TREC Video Retrieval Evaluation*.

Teun A. van Dijk. 1998. Ideology: A Multidisciplinary Approach. *Sage Publications*.

Xiao Wu, Alexander G. Hauptmann, and Chong-Wah Ngo. 2007. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In *Proceedings of the 15th International Conference on Multimedia*, pages 168–177.

Yun Zhai and Mubarak Shah. 2005. Tracking news stories across different sources. In *Proceedings of the 13th International Conference on Multimedia*.

Dong-Qing Zhang, Ching-Yung Lin, Shi-Fu Chang, and John R. Smith. 2004. Semantic video clustering across sources using bipartite spectral clustering. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME)*.

Text Mining Support for Semantic Indexing and Analysis of A/V Streams

Jan Nemrava^{1,2}, Paul Buitelaar², Vojtěch Svátek¹, Thierry Declerck²

Department of Information and Knowledge Engineering
University of Economics
Prague, Czech Republic

Language Technology Lab & Competence Center Semantic
Web, DFKI (GmbH)
Saarbrücken, Germany

E-mail: {nemrava,svatek}@vse.cz, {paulb,declerck}@dfki.de

Abstract

The work described here concerns the use of complementary resources in sports video analysis; soccer in our case. Structured web data such as match tables with teams, player names, score goals, substitutions, etc. and multiple, unstructured, textual web data sources (minute-by-minute match reports) are processed with an ontology-based information extraction tool to extract and annotate events and entities according to the SmartWeb soccer ontology. Through the temporal alignment of the primary A/V data (soccer videos) with the textual and structured complementary resources, these extracted and semantically organized events can be used as indicators for video segment extraction and semantic classification, i.e. occurrences of particular events in the complementary resources can be used to classify the corresponding video segment, enabling semantic indexing and retrieval of soccer videos.

1. Introduction

We present an experiment in the use of complementary resources for the semantic indexing and analysis of audio/visual (A/V) streams, i.e. in the domain chosen (soccer matches) this concerns structured web data (match tables with teams, player names, score goals, substitutions, etc.) and unstructured, textual web data (minute-by-minute match reports). Events extracted from these resources are marked up with semantic classes derived from an ontology on soccer by use of an information extraction system. Through the temporal alignment of the primary video data (soccer match videos) with the textual and structured complementary resources, these extracted and semantically organized events can be used as indicators for video segment extraction and semantic classification, i.e. the occurrence of a 'Header' event in the complementary resources will be used to classify the corresponding video segment accordingly.

This information can then be used for semantic analysis, indexing and retrieval of soccer videos, but also for the selection of A/V features (motion, audio-pitch, field-line, close-up, ...) for specific soccer event types, e.g. a CornerKick event will have a specific value for the field-line feature (EndLine), a ScoreGoal event will have a high value for the audio-pitch feature, etc. As such identification of characteristic features is based on textual evidence we call this 'cross-media feature selection and extraction'.

The remainder of this paper is organized as follows. In section 2 we will discuss the nature and potential use of complementary resources in video analysis. In section 3 we present the experiment we did on using complementary resources in the analysis and semantic annotation of soccer match videos. In section 4 we discuss our approach to the extraction of 'cross-media features' and finally in section 5 we draw some conclusions of our work and look forward to future work.

2. Resources Complementary to A/V streams

Despite the advances in content-based video analysis techniques, the quality of video analysis, indexing and retrieval would strongly benefit from the exploitation of related (complementary) textual resources, especially if these are endowed with temporal references. Good examples can be found in the sports domain. Current research in sports video analysis focuses on event recognition and classification based on the extraction of low-level features and is limited to a very small number of different event types, e.g. 'scoring-event'. On the other hand, complementary resources can serve as a valuable source for a more fine-grained event recognition and classification.

When describing complementary resources we distinguish between two different kinds of information sources according to their direct vs. indirect connection to the video material. Primary complementary resources include such information that is directly attached to the media - namely overlay texts, audio track and spoken commentaries. Secondary complementary resources include information that is independent from the media itself but related to its content - it must be identified and processed first. The next two sections describe each of these in more detail.

2.1 Primary Complementary Resources

Although primary complementary resources are not the main focus of our current research and remain more in the field of low-level analysis, we consider them as a valuable source of relevant information. Apart from the audio track containing spoken commentaries we can make use of overlay text that is present in the video picture. The audio track of sports events is however unfortunately known for a very high Word Error Rate on automatic speech recognition (Sturm et al., 2003), even when dealing with a limited vocabulary such as player names and likely events. We decided therefore not to use the audio track

information in our research. The overlay text (a typical example is the time counter in sport events reporting as shown in Figure 1 below) instead provides us with very important information about the time offset between the video file time and the real match time. This information is crucial for the alignment of events extracted from complementary text resources with the low-level video analysis results.

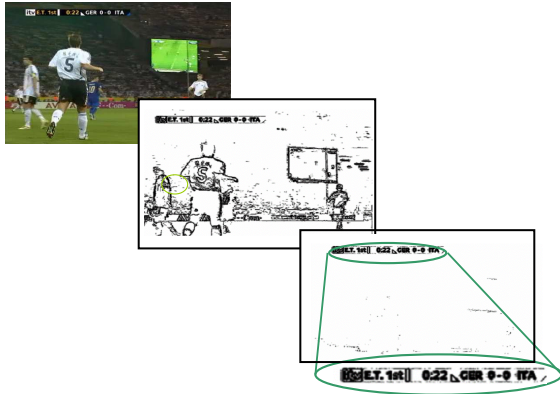


Figure 1: Primary Complementary Resources Example

2.2 Secondary Complementary Resources

The focus of our work is on the use of secondary complementary resources that come in the form of semi-structured tables, containing the summary of statistical, numerical and categorical data connected with events covered by video broadcasts (such as soccer matches) and in the form of unstructured textual reports containing detailed descriptions about particular events covered by video broadcasts including time point information.

Semi-structured as well as unstructured match reports can be readily obtained from web sources and information can be extracted by use of wrappers based on regular expressions in the case of semi-structured tables or of more sophisticated techniques that involve NLP-based information extraction in the case of unstructured text reports (Nemrava et al., 2007).

For our purposes we used semi-structured match tables as well as so-called minute-by-minute match reports, which combine unstructured text information on typical events that are not covered by the tabular match reports with a level of temporal structure through time points, i.e. indication of minute in the match.

3. Semantic Indexing of A/V Streams with Complementary Resources

Major sports events, such as the FIFA Soccer World Cup Tournament that was held in Germany in 2006, provide a range of readily-available resources, ranging from A/V material broadcasted by television or Internet, semi-structured data in the form of tables on web sites, to textual summaries and other match reports. For the research reported here, we used a data set of original videos of television broadcasted matches as primary data,

enriched with complementary information that we extracted from web tables (based on the ‘SmartWeb Data Set’ described below) and textual minute-by-minute reports.

The video material was analyzed independently from the research described here (Sadlier et al. 2005). The analysis results are simply taken as input for our research and consist of video segmentation, with each segment defined by a set of feature detectors, i.e. Crowd detection, Speech-Band Audio Activity, On-Screen Graphics, Scoreboard Presence/Absence Tracking, Motion activity measure, Field Line (for a more extensive discussion see below).

The SmartWeb Data Set¹⁶ is an experimental data set for ontology-based information extraction and ontology learning from text. The data set consists of a soccer ontology, a corpus of semi-structured and textual match reports and a knowledge base of automatically extracted events and entities.

Minute-by-minute reports are usually published at soccer web sites and enable people to “watch” the game in textual form on the web. These reports provide valuable information including the exact time point when each event happened. Combining several of these reports will increase the coverage of events. We therefore identified and collected minute-by-minute reports from the following web sites: ARD, bild.de, LigaLive (in German) and Guardian, DW-World, DFB.de (in English).



Figure 2: Semantic Indexing Demo

By use of the information extraction system SProUT (Drozdzyński et al, 2004) in combination with the SmartWeb soccer ontology (D. Oberle et al, 2007) we were able to derive a domain knowledge base from these resources, containing information about players (a list of players names, their numbers, substitutions etc.), the

¹⁶ http://www.dfki.de/sw-lt/olp2_dataset/

match metadata (basic information about the game can contain information such as date, place, referee name, attendance, time synchronization information) and events (score goals, penalties, headers, etc.).

Obviously, such extracted information can be used to build up a semantic index of players and events in the match. Figure 2 depicts an example application of such semantic indexing implemented with SMIL¹⁷. Various extracted information is aggregated and displayed along with the match video (A/V stream of a television broadcast), providing the user with direct access to events and entities occurring in the selected minute, while also enabling non-linear browsing through the match video.

4. Cross-Media Feature Extraction

Apart from the indexing and retrieval, information extracted from the complementary resources can be used also for the selection of A/V features specific for particular soccer event types. As such identification of characteristic features is based on textual evidence we call this 'cross-media feature selection and extraction'. Using machine learning techniques we try to determine discriminative features of selected football event types and build classifiers assigning the appropriate event type to segments of A/V streams. These classifiers will allow creating a permanent connection between the textual information and the A/V analysis. We test whether the A/V detectors themselves are able to classify events of a certain kind. The following events were selected: foul, free kick, header, shot on goal, corner kick and goal. These events are all of different importance, as reflected also in the A/V streams by the time allocated to replays, crowd reaction, interruption etc.

```

<event_entry>
  → <event_ID>49</event_ID>
  → <from_time>00:12:07:02</from_time>
  → <to_time>00:12:10:11</to_time>
  → <event_type>foul</event_type>
  → <player_1>Campbell</player_1>
  → <team_player_1>England</team_player_1>
  → <player_2>Jancker</player_2>
  → <team_player_2>Germany</team_player_2>
  → <location>ownside</location>
  → <score>0:0</score>
</event_entry>

```

Figure 3: Textual Annotation Example

Data: We used two soccer matches from the Euro Cup 2000, one as training and the other as testing data. We used this data because these matches contained very detailed manual annotation (see Figure 3) created in the context of the MUMIS¹⁸ project. Table 1 has the statistics of selected events. Unfortunately for the goal and corner kick event types the number of instances was insufficient for the experiment and we left them out.

	Match 1 - Training Data -	Match 2 - Test Data -
<i>Foul</i>	31	28
<i>Free kick</i>	18	14
<i>Header</i>	27	22
<i>Shot On Goal</i>	8	17
<i>Corner kick</i>	3	8
<i>Goal</i>	7	1

Table 4: Training vs. Test data

We first aimed at creating a binary classifier for every event type predicting whether the given video segment falls into a particular event type or not, rather than trying to build up one classifier over all event types. In other words, we wanted to know if a particular video segment is for example a foul or not. We later extended the classifier to a ternary classifier aiming at two event types predictions (fouls and shot on goals).

Creating derived values: Two problems occurred when we tried to build up a classifier for soccer events based on the A/V analysis. The first limitation is the generality of video detectors and their low number and the second is the fact that each second (or other time window) of the video analysis will be treated individually without regard to the previous and the next values (and thus behavior in time) of the detectors. We tried to overcome this by adding derived detectors describing the previous and the next values of the detectors in the same time range as the event instance itself (usually 3-5 seconds). We believe that this can help the machine learning algorithms to make a clearer distinction between the different event types. After this preprocessing we had 15 detectors in total. Basic ones are crowd, audio pitch, motion level and close-up detectors, derived ones are the previous and the following average values for each detector and the remaining three denote the proportion between the end-zone, middle zone and other zone of the soccer field based on the field line orientation within the video segment.

Train and test: For the given event type, every event element in the textual annotation file was associated with the appropriate video segment and its A/V analysis for the two matches. These data were labeled as training/testing data. The negative instances (i.e. non-event instances) were created by selecting segments of the A/V streams where none of the selected events occurred.

Building up a model: Decision trees provided the best performance over the given dataset. Table 2 shows the results from the experiment. The first 4 rows are the binary classifier and the results while the last two rows present results from the ternary classifier predicting three classes (2 event types and other)

¹⁷ <http://www.w3.org/TR/REC-smil/>

¹⁸ <http://lands.let.kun.nl/TSPublic/MUMIS/>

binary classifier	total (positive + negative)			event (positive instances)			statistics for event type		
	instances	correct	incorrect	instances	correct	incorrect	Precision	Recall	F-Measure
foul	56	44	12	28	17	11	0,94	0,61	0,74
freekick	42	28	14	14	10	4	0,50	0,71	0,59
header	50	35	15	22	16	6	0,64	0,73	0,68
shot on goal	45	29	16	17	9	8	0,53	0,53	0,53
ternary classifier									
shot on goal	74	45	29	17	3	14	1,00	0,18	0,30
foul				28	17	11	0,63	0,61	0,62

Table 2: Results table

Apart from the classifier we wanted to test¹⁹ whether we can identify A/V detectors that are more discriminative than others for particular event types:

	crowd			audio			motion			closeup			field line		
	P	C	N	P	C	N	P	C	N	P	C	N	M	E	O
Foul	x			x					x				x		
Free kick		x			x								x		x
Header			x					x			x		x		x
Shot on goal					x				x			x	x		x
Foul + shot on goal					x				x			x	x		x

Table 3: Feature Selection

(P, C, N – Previous, Current, Next; M, E, O – middle, end, other)

The results in Table 3 show that different detectors are important for different event types. This potentially allows detecting instances of event types based on observing only those detectors that are discriminative for them (this assumption is also used by the decision tree algorithm). However, a combination of several event types would lead to a conjunction of these discriminative features and would become too general.

5. Related work

There are several ongoing activities dealing with multimodal analysis and mapping across different resources. Very interesting work has been done by Xu (2004), also in the soccer domain. They also proposed a scalable framework that utilizes both internal AV features and external knowledge sources to detect events and identify their boundaries in full-length match videos. Besides detecting events, they focused on discovering detailed semantics and performing question answering. The difference was in the amount of textual sources they used and the number of features in the video analysis. Another related work is SportsAnno (Lanagan, Smeaton, 2007), a video browsing system allowing users to read match reports taken whilst viewing the match video associated with the reports. The main difference is that they used videos summarizes and present free text information without any text processing or inf. extraction. The added value is a possibility for users to add comments as the basis for discussion and searching between all the users of the system. Bertini et al. (2006) used a multimedia ontology and MOM (Multimedia Ontology Manager) to automatically annotate manually pre-selected video clips. They also generated automatic clip subtitles.

¹⁹ In this step we used the following attribute selectors from Weka Machine Learning Tool: CfsSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval

The EU IST Project BOEMIE (Castano, 2007) focuses on the use of multimedia analysis results for population and enrichment of ontologies, in the athletics domain. Most published results of the project deal with still images.

6. Conclusions and Future Work

We presented an approach to the use of resources that are complementary to A/V streams, such as videos of football matches, for the semantic indexing of such streams. We further presented an experiment with event detection based on general A/V detectors supported by textual annotation. We showed that such event-detection based on general detectors can work as a binary classifier quite satisfactorily, but when trained to provide classification for more classes performs significantly worse. Using classifiers similar to those we have tested together with complementary textual minute-by-minute information (providing minute-based rough estimates where a particular event occurred) can help in refining the video indexing and retrieval.

7. Acknowledgements

This research was supported by the European Commission under contract FP6-027026 for the K-Space project. We thank David Sadlier and Noel O'Connor (DCU, Ireland) for providing the A/V data and analysis results.

8. References

- Bertini, M., Del Bimbo, A., and Torniai, C.: Automatic annotation and semantic retrieval of video sequences using multimedia ontologies. MULTIMEDIA '06. ACM, New York, NY,
- Castano S., Espinosa S., Ferrara A., Karkaletsis V., Kaya A., Melzer S., Moller R., Montanelli S., Petasis G.: Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology. In Proc. of International Workshop on Ontology Dynamics (IWOD) ESWC 2007 Workshop, Innsbruck, Austria
- W. Drozdowski, Krieger H.-U., Piskorski J., Schafer U., Xu F.: Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications. In KI 1/2004.
- Lanagan J. and Smeaton A.F. : SportsAnno : What do you think?, RIAO 2007 - Large-Scale Semantic Access to Content, Pittsburgh, PA, USA, 30 May - 1 June 2007.
- Nemrava J., Svatek V., Buitelaar P., Declerck T., Sadlier D., Cobet A., Zeiner H., Petrak J.: Architecture for mapping between results of video analysis and complementary resource analysis., K-Space Public Deliverable 5.10
- Nemrava J., Buitelaar P., Simou N., Sadlier D., Svatek V., Declerck T., Cobet A., Sikora T., O'Connor N., Tzouvaras V., Zeiner H., Petrak J.: An Architecture for Mining Resources Complementary to Audio-Visual Streams. In: Proc. of the KAMC (Knowledge Acquisition from Multimedia Content) workshop at SAMT07, Italy, Dec. 2007.
- Oberle D., Ankolekar A., Hitzler P., Cimiano P., Schmidt

- C., Weiten M., Loos B., Porzel R., Zorn H.-P., Micelli V., Sintek M., Kiesel M., Mougouie B., Vembu S., Baumann S., Romanelli M., Buitelaar P., Engel R., Sonntag D., Reithinger N., Burkhardt F., Zhou J.: DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology) *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 5 (2007) 156-174.
- Sadlier D., O'Connor N.: Event Detection in Field Sports Video using Audio-Visual Features and a Support Vector Machine. *IEEE Transactions on Circuits and Systems for Video Technology*, Oct 2005
- Sturm J., J. Kessens, M. Wester, F. de Wet, E. Sanders, H. Strik. (2003) Automatic Transcription of Football Commentaries in the MUMIS Project. In *EUROSPEECH-2003*, p 1853-1856.
- Xu H., Chua T.: The fusion of audio-visual features and external knowledge for event detection in team sports video. In *Proceedings of the 6th ACM SIGMM Workshop on Multimedia information Retrieval*, 2004

Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access

Judith Klavans^{1,2}, Carolyn Sheffield², Eileen Abels³, Joan Beaudoin³, Laura Jenemann, Jimmy Lin^{1,2}, Tom Lippincott⁴, Rebecca Passonneau⁴, Tandeep Sidhu¹, Dagobert Soergel², Tae Yano⁵

¹Laboratory for Computational Linguistics and Information Processing (CLIP)

at the Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD

²The iSchool at the University of Maryland, College Park, MD

³College of Information Science and Technology, Drexel University, Philadelphia PA

⁴Center for Computational Learning Systems, Columbia University, New York NY

⁵Department of Computer Science, Carnegie Mellon University, Pittsburgh PA

E-mail: jklavans@umd.edu, eileen.abels@ischool.drexel.edu, jeb56@drexel.edu, lj27@drexel.edu, jimmylin@umd.edu, tom@cs.columbia.edu, becky@ccls.columbia.edu, csheffie@umd.edu, tsidhu@umd.edu, dsoergel@umd.edu, taey@cs.cmu.edu

Abstract

We present a system which applies text mining using computational linguistic techniques to automatically extract, categorize, disambiguate and filter metadata for image access. Candidate subject terms are identified through standard approaches; novel semantic categorization using machine learning and disambiguation using both WordNet and a domain specific thesaurus are applied. The resulting metadata can be manually edited by image catalogers or filtered by semi-automatic rules. We describe the implementation of this workbench created for, and evaluated by, image catalogers. We discuss the system's current functionality, developed under the Computational Linguistics for Metadata Building (CLiMB) research project. The CLiMB Toolkit has been tested with several collections, including: Art Images for College Teaching (AICT), ARTStor, the National Gallery of Art (NGA), the Senate Museum, and from collaborative projects such as the Landscape Architecture Image Resource (LAIR) and the field guides of the Vernacular Architecture Group (VAG).

1. Project Goals

Creating access to ever-growing collections of digital images in scholarly environments has become increasingly difficult. Studies indicate that current cataloging practices are insufficient for accommodating this volume of visual materials, particularly for diverse user needs. The goal of the CLiMB project is to leverage text already written about images for automatically identifying, categorizing, filtering and selecting high quality descriptive metadata for image access.

Typically, in libraries and museums, cataloging is performed manually with minimal tombstone cataloging, i.e. the basic set of information (e.g. name of work, creator, date). However, what is usually lacking are rich descriptive terms (e.g. for Picasso's *Guernica*, "screaming horse", "the frozen women", "fauns" and "minotaurs")²⁰. In addition, many legacy records lack subject entries altogether. The literature on end users' image searching practices, though sparse, indicates that this level of subject description may be insufficient for some user groups, including both general users and domain experts with knowledge of specialized vocabularies. Furthermore, the lack of subject-oriented description precludes searching and image analysis across topic area (e.g. searching for works with "minotaurs" as a theme).

Our hypothesis is that automatic and semi-automatic techniques may help fill the existing metadata gap by

facilitating the assignment of subject terms. In particular, we are interested in the impact of computational linguistic technologies in extracting relevant access points from pre-selected texts. The CLiMB Toolkit applies Natural Language Processing (NLP), categorization, and disambiguation techniques over texts about images to identify, filter, and normalize high-quality subject metadata

2. The CLiMB Toolkit

Figure 2 shows a screen shot of the CLiMB Toolkit user interface for an image and text from the National Gallery of Art online collection²¹. Note that the center top panel contains the image, so catalogers can examine items as they work. The center panel contains the input text, with proper and common nouns highlighted. Terms under consideration are displayed below the full text with thesaural information accessible in the right-hand panel. Under this is the term the user has selected for consideration. The right-hand panel gives thesaural information. For normalizing terms, we use the Getty Vocabularies²²: the Art and Architecture Thesaurus (AAT), the Thesaurus for Geographic Names (TGN), and the Union List of Artist Names (ULAN). In this example, two senses for the word "landscape" are displayed on the right. Note that the top portion of the panel displays possible matches in the AAT, followed by the middle portion which shows the chosen definition for the selected term, and finally, the bottom panel in which the entire hierarchy is displayed for the user to view and used to

²⁰ Taken from the exhibition notes from the Picasso exhibit at the National Gallery of Victoria, published by www.thornton.com

²¹ www.nga.gov

²² http://www.getty.edu/research/conducting_research/vocabularies/

identify any related terms.

To extract terms from these relevant segments, we use off-the-shelf software to perform traditional NLP techniques. In the current Toolkit, the Stanford tagger (Toutanova and Manning, 2000; Toutanova et al., 2003) is used since it is Java compliant and currently outperforms other taggers. We have used the open source Lucene toolbox to index. Internally developed noun phrase and proper noun identification rules have been applied. As part of categorization, we have applied a machine learning technique trained over text in the art and architecture domain to select a functional semantic category (Passonneau, 2008). Finally, we explore several disambiguation techniques, which we continue to refine and test with our user groups (Sidhu, 2007). Finally, candidate terms are proposed to catalogers for selection and export into an image database.

Currently, CLiMB focuses on nouns and noun phrases. Recent literature on image indexing indicates, however, that other parts of speech may be valuable in retrieving images. In a study of image professionals, (graphic designers, advertising staff, etc.), Jorgensen (2005) found that “while nouns account for the largest percentage of term type in image searches (just over 50%), adjectives account for 18% of the total term usage, verbs 10%, proper nouns 5%, concept 8%, byline 2%, visual content 2%, and date 1%. Of course, these results are highly dependent on the users and their image needs, but it does give some indication of the relative importances of the term types being searched.”

3. Related Research

Broad domain users (as opposed to specialists) require access using broader non-specialist terms. Choi and Rasmussen (2003) studied the image-searching behaviors of faculty and graduate students in the domain of American history and found that generalists submitted more subject-oriented queries than known author and title searches. Currently, much cataloging is geared towards the specialist. On the other end of the spectrum is pure indexing of textual material in the physical domain of an image, such as that done by google (Palmer n.d.). Although such approaches are valuable for initial image access, the resulting high recall can make for a frustrating browsing experience for the end user.

On the other hand, the subjective nature of images inherently complicates the generation of accurate and thorough descriptions. Berinstein (1999) points out that even the guidelines provided by the Shatford-Panofsky matrix on what to describe are fluid and may be difficult to apply. Shatford (1994), building on Panofsky (1962), proposed a method for identifying image attributes, which includes analysis of the generic and specific events, objects, and names that a picture is “of” and the more abstract symbols and moods that a picture is “about”. Panofsky describes the pre-iconographic, iconographic, and iconologic levels of meaning found in Renaissance art images. Shatford's generic and specific levels correspond to Panofsky's pre-iconographic and iconographic levels, respectively, and encompass the more objective and straightforward subject matter depicted in an image. The

iconologic level (Shatford's about) addresses the more symbolic, interpretive, subjective meanings of an image. To aid user access, catalogers are encouraged to consider both general and specific terms for describing the objective content of an image as well as to include the more subjective iconologic, symbolic, or interpretive meanings. Iconologic terms may be the most difficult for catalogers to assign but occur often in texts describing images.

4. Current Cataloging Approaches

In the CLiMB workflow studies, we examined existing cataloging practices and gathered cataloger perspectives on current challenges in image indexing. Understanding the component processes in current practice has enabled the development of the CLiMB workbench to be easily integrated into existing standards, systems, and practices. Furthermore, by determining which challenges are general to the field and which arise in conjunction with specific collections, we were able to identify additional needs which our research may address. In architecture collections, for example, text may describe a building or architectural site as a whole while the corresponding image typically provides only a detailed view of the work. Part-whole relationships such as these present specific linguistic challenges for associating segments of text with one or more images. This research is not the topic of this paper, and will be described in a forthcoming article.

5. CLiMB Architecture: Systems and Methods

The CLiMB architecture is shown in Figure 1. The data flow for CLiMB starts at the upper left which shows the input to the system:

1. an image,
2. minimal metadata (e.g. image, name, creator)
3. text.

This input is pre-processed, using external technologies, to identify coherent segments of text and associate those segments with relevant images. Input texts are marked up using TEI lite (Text Encoding Initiative) to identify topical divisions (chapters, sections, etc.). These divisions, or segments, are then mapped to corresponding images through the identification of plate and figure numbers. For art historical survey texts, such as Jansen (2004) and Gardner (2001), the automation of text-image association produces reliable results. CLiMB has investigated the application of linguistic technologies to semi-automatically classify, or categorize, text segments according to their semantic relationship to the image(s) which they describe Passonneau, et al (2007).

Through our partnership with the Getty Research Institute, we have been given access to three resources:

- The Art & Architecture Thesaurus (AAT), a structured vocabulary for describing art objects, architecture, and other cultural or archival materials. The AAT's structure is comprised of seven major facets (Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials, and Objects) from which

multiple hierarchies descend. In total, AAT has 31,000 such records. Within the AAT, there are 1,400 homonyms, i.e., terms that can lead to several AAT records that may have multiple meanings only one of which may apply in a given context.

- The Union List of Artist Names (ULAN), a name authority that includes the given names of artists, as well as any known pseudonyms, variant spellings, and name changes (e.g., married names). The structure of this resource is similar to the Agents facet of the AAT in that it contains Person and Corporate Body as its primary facets.
- The Thesaurus of Geographic Names (TGN), an authority for place names, including place names as they appear in English as well as in other languages, historical names, and names in natural order and inverted order.

These vocabularies are well-established and widely-used multi-faceted thesauri for the cataloging and indexing of art, architecture, artifactual, and archival materials. Each of these resources specifies which variation of a given concept or name is the preferred term, enabling consistent cataloging across collections. We have utilized these resources to link terms derived from testbed texts to standardized, controlled terms, thus helping users expand their information space. The Getty resources are used to select the particular homograph of a term.

5.1 Disambiguation

We have tested three approaches to disambiguation in our domain, using the AAT as our baseline thesaurus (Sidhu, 2007). However, it is clear that we need to utilize additional terminological resources since many common terms—and senses of ambiguous terms—are missing from the specialist thesaurus. The challenge of using domain-specific vocabularies combined with general vocabularies, and the impact on disambiguation, is a little-studied topic. We have observed that terms with many senses in the AAT may have just one sense in a general dictionary, and that some terms with many senses in a general resource are simply missing altogether in the AAT. The impact of these observations on disambiguation has yet to be established.

In order to test our disambiguation technique, we first annotated a text to use for evaluation. Following standard procedure in word sense disambiguation tasks (Palmer et al., 2006), two labelers manually mapped 601 subject terms to the AAT. Inter-annotator agreement for this task was encouragingly high, at 91%, providing a notional upper bound for automatic system performance (Gale et al., 1992). We have used SenseRelate (Banerjee and Pederson, 2003; Patwardhan et al., 2003) for disambiguating AAT senses. SenseRelate uses word sense definitions from WordNet 2.1, a large lexical database of English nouns, verbs, adjectives, and adverbs.²³

Results from our evaluations (discussed in Sidhu et al, 2007) show that mapping to WordNet first and then to the

AAT causes errors. As a general resource, WordNet is domain independent and thus offers wider, more comprehensive coverage. However, the lack of domain specificity also creates overhead as there are many irrelevant senses to choose from and the correct sense needed for art and architecture discourse may not be available. Similarly, Iyer and Keefe (2004) report on an exploratory study on the use of WordNet to clarify concepts for searching architectural visual resources. Twenty participants were shown images which they were asked to locate using natural language or WordNet terms. Although 70% of participants stated that WordNet clarified the terms or the images, 30% reported problems with conceptualizing the image, and 55% had terminology problems. To address these types of problems, we are exploring the option of re-implementing concepts behind SenseRelate to directly map terms to the AAT. Additionally, in Future Work we will test approaches for employing hybrid techniques (including machine learning) for disambiguation. This will enable us to explore the trade-off in precision between different configurations of resource calling.

5.1.1. Catalog Record Creation: Select

As shown in Figure 2, a cataloger is presented with the image to be cataloged, the text segment associated with the image, and a number of index terms suggested by the Toolkit. The user decides which of the terms proposed by the CLiMB system should be included in the image's record.

5.2 Testbed Collections

We are currently working with five image-text sets and one image collection for which we are conducting experiments with dispersed texts located online. Table 1 illustrates the relationship between the associated texts and the image collections which we use to test our system.

Feedback from catalogers indicates that one thesaural resource is insufficient for cataloging a range of art historical and architecture images. The Getty resources are extensive but, as with any resource, are not entirely comprehensive. Our goal is to expand our capabilities for disambiguating domain-specific terminology by cross-searching multiple, established thesauri in the art and architecture domain. Resources currently under consideration include Iconclass²⁴ and the Library of Congress' Thesaurus for Graphic Materials (TGM) I and II^{25,26}.

6. Conclusion and Future Work

The CLiMB project techniques exceed simple keyword extraction and indexing by:

- applying novel semantic categorization to text segments,
- identifying and filtering linguistically coherent phrases,
- associating terms with a thesaurus, and

²⁴ <http://www.iconclass.nl/>

²⁵ <http://www.loc.gov/rr/print/tgm1/>

²⁶ <http://www.loc.gov/rr/print/tgm2/>

²³ <http://wordnet.princeton.edu/>

- applying disambiguation algorithms to these terms.

Although each of these techniques has been used in other projects, they have not been combined and tested in the art and architecture domains for improving digital library access. Our future work will consist of three foci:

- Integration of functional semantic categorization with disambiguation
- Improvement of disambiguation
- Testing the system and its components with users to drive improvements

We also hope to incorporate the output of CLiMB text data mining with a social tagging approach to image labeling, such as that of *steve.museum* to examine terminological comparisons and their impact on image access.

Acknowledgements

We acknowledge input on the project from Dr. Murtha Baca of the Getty Vocabularies Institute. We also appreciate ongoing input from both our internal and external advisory boards, the members for which are listed on our web pages. Finally, many users whom we cannot name have helped in evaluations along the way. The project has been funded by the Andrew W. Mellon Foundation to the Center for Research on Information Access at Columbia University. Later support was provided to the College of Information Studies and the University of Maryland Institute for Advanced Computer Science at the University of Maryland, College Park, MD.

References

Banerjee, S., Pedersen, T. (2003) Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pp. 805–810.

Berinstein, P. (1999). Do you see what I see? Image indexing principles for the rest of us. Online, 23(2), 85-87.

Choi, Y., Rasmussen, E. M. (2003). Searching for images: The analysis of users' queries for image retrieval in American History. *Journal of the American Society for Information Science and Technology*, 54.6: 498-511.

Gale, W., K. W. Church, and D. Yarowsky. (1992) "Estimating upper and lower bounds on the performance of word-sense disambiguation programs." Proceedings of the 30th conference on Association for Computational Linguistics. pp. 249-256.

Gardner, Helen. (2001) *Art through the Ages*. 11th edition. Edited by Fred S. Kleiner, Christin J. Mamiya, Richard G. Tansey. Harcourt College Publishers, Fort Worth, Texas.

Iyer, H., & Keefe, J. M. (2004). WordNet and keyword searching in art and architectural image databases. *VRA Bulletin*, 30, pp. 22-27.

Janson, H. W. (2004) *History of Art: the Western*

tradition. Revised 6th edition. Upper Saddle River, NJ: Pearson/Prentice-Hall.

Jorgensen Corinne, and Peter Jorgensen. (2005) Image querying by image professionals. *Journal of the American Society for Information Science and Technology*. Hoboken: Oct 2005. Vol. 56, Iss. 12; pg. 1346.

Klavans, Judith L. (2006) Computational Linguistics for Metadata Building (CLiMB). In Proceedings of the OntoImage Workshop, G. Greffenstette, ed. Language Resources and Evaluation Conference (LREC), Genova, Italy.

Palmer, Justin (n.d.) Optimizing Your Images for Google Image Search - 4 Image SEO Tips. <http://ezinearticles.com>.

Palmer, M., Ng, H.T., Dang, H.T. (2006) Evaluation. In: Edmonds, P., Agirre, E. (eds.): *Word Sense Disambiguation: Algorithms, Applications, and Trends*. Text, Speech, and Language Technology Series, Kluwer Academic Publishers, Netherlands.

Panofsky, E. (1962). *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. Harper & Row, New York .

Passonneau Rebecca, Tae Yano, Judith Klavans, Rachael Bradley, Carolyn Sheffield, Eileen Abels, Laura Jenemann (2007) Selecting and Categorizing Textual Descriptions of Images in the Context of an Image Indexer's Toolkit. Technical Report. Columbia University.

Passonneau Rebecca J., Tae Yano, Tom Lippincott, and Judith Klavans. (2008). Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning. International Workshop on Metadata Mining for Image Understanding; VISAPP International Conference on Computer Vision Theory and Applications. MMIU 2008. Funchal, Madeira - Portugal.

Patwardhan, S., Banerjee, S., Pedersen, T. (2003) Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.

Shatford-Layne, S. (1994): "Some issues in the indexing of images." *Journal of the American Society for Information Science* 45.8. pp. 583-588.

Sidhu, T., Klavans, J.L., Lin, J. (2007) Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTech 2007), 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic

Toutanova, K., and C. D. Manning. (2000) Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *EMNLP/VLC 2000*. pp. 63–70.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings

of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1. Edmonton, Canada: Association for Computational Linguistics. pp. 173-180.

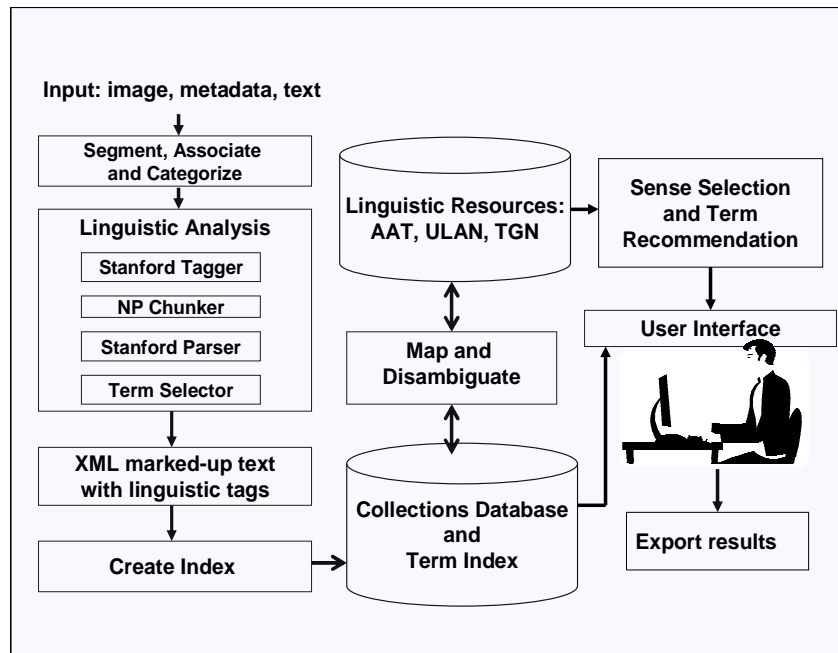


Figure 1: CLiMB Architecture


Image Collection	Text	Image/Text Relationship
National Gallery of Art (NGA) Online Collection	Narratives associated with images on the NGA website	Integrated
U.S. Senate Museum	U.S. Senate Catalogue of Fine Arts	Integrated
The Vernacular Architecture Forum (VAF)	VAF Field Guides:	Integrated
The Society of Architectural Historians (SAH): World Architectural Survey and the American Architectural Survey	<i>Buildings Across Time: An Introduction to World Architecture</i> by Marian Moffett, et al.	External
Landscape Architecture Image Resource (LAIR)	<i>Landscape Design: A Cultural and Architectural History</i> by Elizabeth Barlow Rogers	External
Art History Survey Collection (AHSC), ARTstor	Disparate texts located online	External

Table 1: Sources of Image and Testbed Text Collections

CLiMB Cataloger Workbench: The Martyrdom of Saint Catherine

Export View Help

Image Image Information



Text

As Europeans came to understand the world through printed maps and geographies, landscape emerged as a popular subject. The Italian artist and biographer Giorgio Vasari noted that "there is no cobbler's house without its landscape because one becomes attracted by their pleasant view and the working of depth." Here the torture of Saint Catherine is overwhelmed by the scenery in a way typical of the panoramic "world landscapes" painted by northern artists. Perspective and point of view are manipulated to provide the most information possible: this is a God's-eye view that sees everything simultaneously. Varied terrain and captivating detail compel the viewer to travel across the picture with his eyes. This painting may be the work of Matthys Cock, whose brother Hieronymus was a well-known publisher of prints, including many by Pieter Bruegel. Notice the distinctive mountain crags here and similar ones in other landscapes nearby painted by

Subject Terms

Term Under Consideration

landscape

Terms Assigned

Term	Normalized Term	Subject ID
Patnir	Patnir, Joachim (Patnir, Joachim, Person)	ULAN:500019...
mountain	mountains(landforms by shape or position: upward, landforms by sh...	AAT:300008795
rock	rock(inorganic material, materials by composition, ... Top of the AAT ...	AAT:300011692
panorama	panoramas(visual works by form: image form, visual works by form,...	AAT:300015537

AAT Browser(2) TGN Browser(1) ULAN Browser(9)

Search Term: landscape Show Partial Match

landscapes [Settlements and Landscapes]

landscapes [visual works by subject type]

Selected Record Description

Use for creative works that depict outdoor scenes where the picture is dominated by the configuration, visual and aesthetic, of the land, bodies of water, and natural elements. When the ocean or other large body of water dominates the picture, use "seascapes." For images that are more documentary than creative, prefer "views" or "topographical views." For actual areas of land having certain notable characteristics, use "landscapes (environments)."

Selected Record Hierarchy

- Exchange Media
 - Information Forms
 - Visual Works
 - visual works
 - visual works by form
 - visual works by function
 - visual works by location or context
 - visual works by medium or technique
 - visual works by subject type
 - Buddhas
 - Christmas trees
 - animal paintings
 - capricci
 - cityscapes
 - crosses
 - drapery
 - figures
 - genre
 - history paintings
 - landscapes**

Show Full Display

Legend: Selected Term Terms With Children

Figure 2: CLiMB User Interface for the term "landscape"