The role of context in image interpretation

Dag Elgesem[1] and Joan Nordbotten[1]

[1]Department of information science and media studies, University of Bergen
P.O. Box 7800, 5020 Bergen, Norway

{Dag.Elgesem, Joan.Nordbotten}@infomedia.uib.no

Abstract. With the increasing availability and use of mobile phones with camera functionality, it becomes feasible to use these devices to query image databases for information. However, the resulting photo/image query commonly depicts several objects and can be open to a number of interpretations. This leads to ambiguities when determining the intention of an image used as the basis for an information retrieval query. We suggest that this problem is structurally similar to the problem of how to interpret an ambiguous sentence, and that the task can be modeled in a similar way. Though the role of context is a key factor in the solution of the problem of disambiguation of text, we argue that existing accounts of context do not explain what role context plays in image interpretation. In this paper, we propose a definition of image *context* and then show how the disambiguation of images as queries can be modeled as a game of partial information. On the basis of this, a more precise account of the role of context in image interpretation is proposed.

Introduction

Increasing numbers of digitized image collections are available on the Internet and can be accessed via search engines or specialized image retrieval systems. These systems are also available to anyone possessing a mobile phone with Internet connection and camera/image functionality, making image-based query formulation feasible for a broad user community. Unfortunately, current image retrieval algorithms do not yet have the effectiveness of their counter-part text retrieval algorithms, when measured by the degree of relevance of the result sets for the user.

There are 2 main approaches to image retrieval. The most common approach used for Internet access to image databases is a keyword match based on annotations that have been manually defined for each image. This is a tedious task that may not provide good

descriptors for the information seeker [5]. An alternative approach, called content-based image retrieval, CBIR, uses an input image as the query statement which is matched to the structural characteristics (color, texture and shape) of the stored images. This approach suffers from the gap between the user's understanding of the semantic meaning of the search image and the current inability of the image retrieval algorithms to identify objects within the image and thus recognize its semantic meaning [1, 5]. Our work[1] is focused on improving the quality of image retrieval based on visual (image) queries by automatically extending the context information in the annotations of image collections and by using context information in the interpretation of visual queries. In this paper we will address the second issue.

An initial problem when an image is used as a query is how context information can be used to determine the user's intention in submitting the image query. In comparison to text queries, an image has no regular structure or components with a defined definition and grammar, and thus far fewer constraints on its interpretation than a string of text.

Consider the following scenario, which could be typical for a tourist with a camera phone. Our tourist is walking around in the city and stops in front of an old church that she finds interesting. There is also a group of sculptures clearly visible above the church door. The tourist wants to know more about the church and pulls out her camera phone, takes a picture of the front of the church, including the sculptures above the church door, and sends it to a multi-modal information retrieval system, MIRS [5] for historical information, expecting to get information about the church. The problem, now, is how can the MIRS determine that she in fact wants general information about the whole church, and not specific information about just the sculptures, when both objects are depicted in the photo?

We will address two general issues in the connection with this question. First, we will suggest that the structure of the problem about the interpretation of the user's intention on sending the picture is analogue to the problem of how an ambiguous sentence is interpreted. We will argue this point in more detail below, but state here that context plays a central role in the process of disambiguation of a

[11] Financed by the Norwegian research council, NFR, Project # 176858/S10: Context Aware Image Management http://caim.uib.no/

sentence. This brings us to the second general issue we will address here: what role does context play in the interpretation of an ambiguous sentence, and in the disambiguation of a query in the form of an image?

Let us start with the latter issue. There are a number of definitions of context in the literature (e.g. [2], [3], [6], [8]) and many of them define context in terms of the role the contextual information plays in the interaction between the system and the user. A central example is

Dey's [2] definition of context as

> *any information that can be used to characterize an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves.* (p. 5)

While we agree that Dey's definition is basically correct, it is clear that the reference to what "is considered relevant to the interaction" leaves a lot to be explained. When one considers particular examples of communication it is often easy to point out what information is relevant. But it is important at a theoretical level to explain how and in virtue of what information becomes relevant in a given situation. We suggest that this explanation can be given only by way of an analysis of the structure of communication.

It should be mentioned, however, that Dey does give a partial answer to the question about when information becomes relevant. In his definition of a context aware application he suggests that relevance is relative to the user's task:

> *A system is context-aware when it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task. (p. 5)*

Again, we share the spirit of Dey's definition but claim that it leaves important questions about how information becomes relevant unanswered. Thus, the definition cannot be made operational. Something more is needed to explain how information becomes contextually relevant.

The problem is that the definition seems to treat the user's task as something that is given. But in the cases we are considering, with ambiguous images, the problem is exactly to figure out what the user's task is. Since context clearly has to be involved in the determination of what the user's task is, we cannot look at the user's task to determine

what information is relevant. The system has to use the context to find out what the user's task is. We cannot in general assume that the user's task is known but want to understand how the application can use the context of the communication to determine the right interpretation of the query.

Following the above observations, our suggestion for a definition of context is:

> *Context is the information that must be common knowledge between user and system for communication between them to succeed.*

This definition might not seem radically different from Dey's, but is in fact different in important respects that we will try to make clear through the discussion below.

That this is a reasonable definition of context is one of the points of the paper. The other one, mentioned above, is the suggestion that the problem of determining the user's intention in sending an image to a MIRS is a special case of the problem of disambiguation. We have two arguments for this. The first is that it is intuitively plausible. Consider a person who utters the ambiguous sentence "Every day a man is mugged in Bergen". It seems clear that the most likely interpretation is that "every day there is some man or other who is mugged", rather than "there is one particular man who is mugged every day". How do we know this? Because, first, on the basis of general knowledge about the world. Second, because we can assume that this is common knowledge and that if the speaker had intended the second interpretation he would have used a sentence that was not ambiguous. The point is that the disambiguation cannot happen without bringing in the context. The situation is exactly similar when a statement is made in the form of an image: both the background information about the world and reasoning about what is common knowledge have to be brought to bear in the choice of an interpretation.

Our second argument for the claim that the interpretation of an image is similar to the disambiguation of a sentence is that a model of the latter can be used to analyze the former. In the following we will present and discuss this model of disambiguation.

We are not alone in suggesting that conditions for successful communication are important to the analysis of context and relevance. Mani and Sundaram [6] argue that the key to the understanding of

context is to analyze its role in communication. They define context as a “finite and dynamic set of multi-sensory and inter-related conditions that influences the exchange of messages between two entities in communication.” ([6], p. 340) Our approach to the analysis of communication is however different from theirs because we develop our notion of context on the basis of the concept of a game of partial information. The role of common knowledge in the delineation of the context will thus be made explicit in our approach while this is only implicit in [6].

Parikh’s model of disambiguation

There are many pieces of information that might be relevant to determine the user’s context. There is information about location, general background information, information from analysis of the image, etc. But what information is actually useful in the interpretation of a given image-query? Before we can answer this question, an account is needed of the precondition for the common determination of the meaning of an ambiguous query.

In his book, *The Use of Language*, Parikh [7] develops an account of how two communicating agents achieve understanding of the intended meaning of an ambiguous sentence. To this end he uses the framework of games with incomplete information. Applying his theory to our setting, assume that we have a human user U and an automated system S communicating via photographs. The user sends pictures to the system, and the system tries to determine what informational need the image indicates, and sends relevant information back to the user.

The problem can be modeled as a game with partial information. Assume that U moves first and sends a picture pic1, e.g. a picture of a church, to S, and that pic1 has two interpretations ‘church’ (i.e. the whole church) and ‘sculptures’ (i.e. the sculptures on the church wall). The picture is visible to both actors, and hence common knowledge. Assume, further, that U’s intention in sending the picture to S is to communicate the first interpretation, i.e. that she wants to know more about the whole church. But since S does not have direct access to U’s mind, it has to infer it on the basis of general assumptions about

the situation and information about the context. There are several further aspects of the context that play a role in the making of this inference, as the model will make clear. Let us explain the details with reference to the following figure:

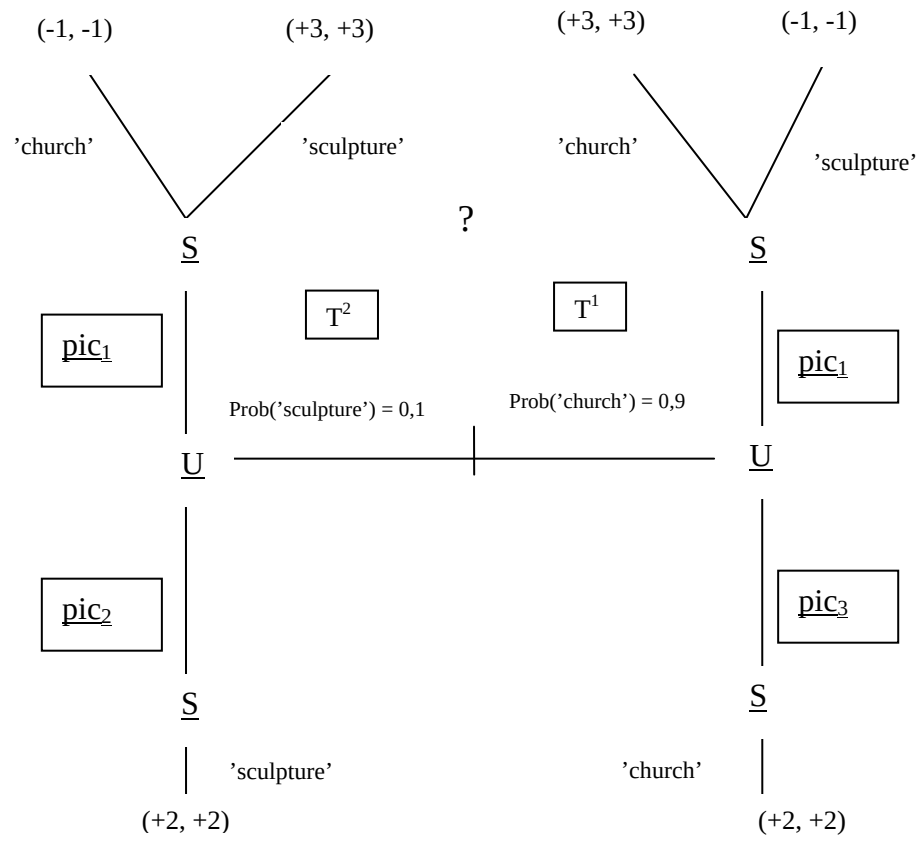

Fig. 1. Image interpretation

By sending the picture, U could either intend to indicate that she is interested in the church or in the sculptures. If the first is the case, U would be in situation T^1. If the intention is to indicate the sculptures, S would be in T^2. The right side of figure is a representation of the situation where U wants to indicate to S that she is interested in the

whole church, i.e. T^1, while the left side represents the situation T^2. We see that, as indicated on the central horizontal line, U is more likely (0,9) to want information about the whole church than the sculptures in particular (0,1). This is assumed to be a fact about U at this point of the interaction. (After she has received general information about the church, the probability that she wants information about the sculptures will perhaps increase.) Note that there is a real chance that she would send pic_1 even when she wanted to know more about only the sculptures (i.e. in T^2), hence the ambiguity.

But even though U of course knows her own intentions, S can observe only the picture (pic_1) and cannot be sure whether it is in situation T^1 or T^2. The problem is, again, how can S rationally be sure of U's intention in this game? For this to be possible, two more elements are needed. The first is that knowledge about alternative ways of depicting the object of interest that are not ambiguous. For example, a close up picture of only the sculptures would unambiguously indicate the sculptures. Similarly, a picture taken from a longer distance of the whole church without any surrounding buildings would unambiguously indicate an interest in the church in general. In the figure above, these alternative ways for U to indicate her intention appear in the lower half of the diagram, and are called pic_2 and pic_3, respectively. We see that these alternative ways of depiction have only one interpretation, and one that unambiguously expresses U's intention.

The second element that is needed for S to be able to solve the problem of determining U's intention, is that the parties have to assign values to the possible outcomes, i.e. that a payoff function is defined. There are two factors that will affect the values of the outcomes: the costs of taking the various pictures, and the utility of the chosen interpretation. It is assumed that it is more costly to take a close up picture of a detail than of the whole wall of the building. And, again, it is assumed that a correct interpretation has a positive value, and that misinterpretation has a negative value, for both the user and the system.

To see how this works, consider first the upper, right-hand part of the figure. Here U sends pic_1 to S with 'church' as the intended interpretation. If, now, S chooses 'church' as the interpretation of the picture, this is a positive outcome for both. On the other hand, if S chooses 'sculptures' as the interpretation, we have a case of miscommunication and this would be a negative outcome for both. (Technically, the values assigned to the outcomes could be different,

but since the parties are cooperating, it is fair to assume that they have the same valuation.) The situation is different in the upper left part of the figure. Here U sends pic_1 to S with 'sculptures' as the intended interpretation. If S here chooses 'church', this would mean a breakdown of communication and thus a negative outcome, while 'sculptures' would be a positive result for both.

Consider now the lower part of the figure. On the right side, i.e. in situation T1, U sends pic_3 which unambiguously indicates to S that 'church' is the intended interpretation. There is only one outcome and this secures a positive outcome. Similarly on the left side, in situation T^2, where pic_2 unambiguously indicates to S that U's intended interpretation is 'sculptures'. We see that the outcomes in this case, even though they are positive, are valued lower than the positive outcomes in the upper part of the diagram.

This brings us to the second element in the solution of the game of disambiguation of the image. In order to achieve this, U and S "need to compare this ambiguous utterance against an unambiguous one, to ensure that it is more efficient". (Parikh in [7], p. 30) The point is that the outcomes are different because it takes more effort and is thus more costly to create an unambiguous picture. For example, U could have moved up closer to the church and focused only the sculptures (i.e. pic_2). Or, again, she could have moved farther away and taken a picture that captured the whole church and without the sculptures clearly visible (pic_3). This would have communicated that the church was the object of interest. But in both cases the communicative success comes with a price: the extra cost involved in taking more precise pictures. These extra costs are the reason the payoffs are lower in these cases.

The assumptions of common knowledge

With these elements in place, it is possible for a rational agent to determine U's intended interpretation. To reach a unique solution to this problem, a number of requirements have to be met.

1) Both of the agents have to be rational, i.e. their preferences are consistent and transitive, and they maximize outcomes.

2) They have to share a system of ways to depict objects, i.e. there is a language of a sort that is common knowledge.

3) There has to be common knowledge about how structures in the pictorial language refer to objects in the real world.

4) In a given situation there has to be common knowledge of what the possible interpretations are, i.e. that the picture in question (pic_1) is ambiguous. Hence, it is common knowledge that S knows that it is in situation T^1 or situation T^2, but not which.

5) There has to be shared knowledge about how relatively likely the various interpretations are. In our example, it has to be assumed to be common knowledge that the first interpretation ('church') is more likely than the second.

6) The values distributed to the various outcomes by the payoff-function also have to be common knowledge. In our case, this also means that it has to be common knowledge that referring to the objects unambiguously takes greater effort than referring to them ambiguously.

These are of course highly non-trivial assumptions, an issue to which we will return to briefly in the conclusion below. But the important theoretical point for now is that on the basis of these assumptions, it can rigorously be shown that a rational sender will choose the signal that is most efficient and that a rational receiver will end up with the intended interpretation. It is not necessary for our purposes to go into the details of the proof that a unique equilibrium exists, which involves both the idea of a Nash-equilibrium and that of a Pareto-dominance between strategies. The interested reader is referred to Parikh's superb exposition [7].

Context and common knowledge

Parikh's model offers a very powerful account of interpretation and disambiguation of sentences in natural language. The model is helpful with regard to the discussion of the definition of context with which

this paper started. The problem with the definition provided by Dey was that it involved essentially the notion of relevance. But this leaves unanswered the question of what information is relevant. But this is exactly what the model discussed above provides. The information that is relevant, and hence makes up the context, is exactly the information that has to be common knowledge in order for communication to succeed.

As the discussion above shows, the framework can be applied to the modeling of the interpretation of images. We believe that even if it abstract, this rational reconstruction of what it is to disambiguate an image is a useful first step in the process of creating a MIRS. It provides a plausible model of what parts of the context that has to be common knowledge in order to secure the communication of the intended interpretation. Hence, the model gives a standard for what the optimal solution to the problem of disambiguation would be and it helps us see where a concrete system is an implementation of a real solution and where we have to make simplifying assumptions.

But we also believe that the framework can in fact provide the basis for the development of a MIRS of the sort indicated in the scenario discussed above. This suggestion gives raise to several questions.

First, images do not depict objects in the same ways as written language does (as established by Goodman in [4]). They do not have syntactic and semantic structures analogous to writing.

A second problem is the amount of world knowledge that is needed to be able to undertake the disambiguation, i.e. to know the set of possible interpretations, their respective probabilities and the valuation of various outcomes.

The third problem is to establish the vast range of common knowledge that is needed in order to solve the problem of disambiguation. This is also a highly non-trivial problem.

A forth problem we will mention is related to the assumption of rationality. It is well known that humans are not perfectly rational in the sense specified by rational choice theory and a context-aware application should be able to take this into account.

Finally all of this information must become common knowledge between the user and the system.

Discussion

We believe that with respect to the class of situations indicated by the scenario above there are practical solutions to all of these problems. First of all, the system we are about to develop is location based in the sense that the images will be tagged with GPS data. It is, of course, possible to automatically recognize objects in images given enough world knowledge in a limited domain. For example, in the CAIM project, mentioned above, we will address this with the use of location data together with image analysis of photographs of a restricted set of historically interesting buildings and objects. Also, an image will contain rich information about the user's point of view with respect to the objects that are the topics of attention. Hence the image that is used as a query will have an indexical function that will limit the set of possible references. In addition, we will be focusing on a very restricted domain where it is possible to provide background information about the kinds of information that are available, i.e. information about objects of historical interest that can be encountered on a walk through the city. Furthermore, the medium itself – the camera phone – put constraint on the range of possible objects. The objects to be communicated about have to be accessible for a photographer. In sum, these constraints will be sufficient to overcome the problems identified above, we believe. This has to be proved in practice, of course, by constructing a working MIRS.

Hence, even if the model cannot be taken directly as a blueprint for the implementation of an application that is able to disambiguate an image, the account still provides important guidance for the development of such a framework. Our claim is that the clearer understanding of the role of context that the game theoretic model provides is a useful basis for the development of context aware image management.

References

[1] Castelli, Vittorio and Lawrence Bergman. 2002. *Image Databases – Search and Retrieval of digital Imagery*. John Wiley.
[2] Dey, Anind K. 2001. Understanding and Using Context. *Personal and Ubiquitous Computing*. **5**:4-7.
[3] Dourish, Paul. 2004. What we talk about when we talk about context. *Personal and Ubiquitous Computing*. **8**:19-30
[4] Goodman, Nelson. 1976. *Languages of Art*. Hackett. Indianapolis.
[5] Lu, Goujun. 1999. *Multimedia Database Management Systems*. Artech House.
[6] Mani, Ankur and Hari Sundram. 2007. Modeling user context with applications to media retrieval. *Multimedia Systems*. **12**:339-353.
[7] Parikh, Prasanth. 2001. *The Use of Language*. CSLI Publication. Stanford, CA.
[8] Spink, Amanda and Charles Cole (eds.). 2005. *New Directions in Cognitive Information Retrieval*. Dordrecht: Springer.