Dialogue-Based Information Retrieval from Images

Pavel Hamřík, Ivan Kopeček, Radek Ošlejšek, and Jaromír Plhák

Faculty of Informatics, Masaryk University Botanicka 68a,
602 00 Brno, Czech Republic
{xhamr,kopecek,oslejsek,xplhak}@fi.muni.cz

Abstract. Our concept of communicative images aims to provide graphical information by means of dialogue interaction, which is suitable for people with various disabilities. Communicative images are graphical objects integrated with a dialogue interface and linked to an associated knowledge database which stores the semantics of the objects depicted. This paper deals with the utilization of formal ontologies for the process of image annotation and dialogue-based investigation in the context of assistive technologies.

Keywords: Ontologies, Picture Semantics, Dialogue Systems.

1 Introduction and Related Work

This paper deals with the utilization of formal ontologies for the process of image annotation and dialogue-based investigation. The role of ontology in relation to information systems is described in [5]. The paper [11] thoroughly describes the process of building user and expert models, as well as the Web Ontology Language and its use. A system for ontology based annotation and indexing biomedical data is studied in [15] and [4]. The paper [14] introduces medical ontology. Paper [17] describes how CHIP and iCITY systems communicate and exchange user data to obtain a more exact view of the users' interests.

A "communicative image", originally introduced and discussed in [9], is a graphical object integrated with a dialogue interface and linked to an associated knowledge database which stores the semantics of the objects depicted.

The interface between natural language and a formalized ontology framework provides an engine that transforms natural language into corresponding formal schemes. Typically, we can restrict ourselves to a small fragment of natural language, so that the engine can be based on relatively simple grammars in combination with the frames technology and standard techniques for misunderstanding solving. For instance, the question *"How far is it from this hotel to the nearest beach?"* is resolved using the template *"How far is it from SLOT1 to SLOT2?"*. The system expects both the SLOT1 and the SLOT2 to be filled by the specific entries from the ontology. Main principles and details of the dialogue management have been discussed in [7].

K. Miesenberger et al. (Eds.): ICCHP 2014, Part I, LNCS 8547, pp. 85–92, 2014.

A single communicative image consists of three data structures: (a) graphical data, (b) identification of objects in the image and (c) their semantic data, i.e. picture annotations and their associated knowledge base. In our approach we exploit the SVG format [3] to encode all these data structures in a single file. The semantics are encoded as OWL ontologies [10]. Formal ontologies present the key feature of communicative images because they define and structure the vocabulary that is shared across different pictures with similar content. Moreover, they provide a suitable formalism for information retrieval and machine-generated dialogues.

Because users can communicate with standard images, either on the Internet or locally, it is necessary to provide an automatic conversion of common images into a communicative form. The necessary infrastructure is cloud-based, with thin clients and a shared server. The role of the clients is to handle a user's interactions with the image, whether mouse clicking, keyboard typing or voice recognition, and to redirect these interactions to a remote server which contains the core application logic. It is responsible for semantic analysis, reasoning, knowledge storing, management and sharing, and dialogue management.

At the beginning of the communication the client sends the original image to the server. The server acquires as much information about the image as possible using image archives, auto-detection and image recognition algorithms, EXIF data extraction, inspection of shared knowledge database, etc. Then it drives the communication, generates dialogues, searches and filters semantic data stored in the knowledge database, learns from the dialogue and updates the knowledge database.

The cloud-based server approach allows the development of variable thin clients. These clients can be either specialized such as those adapted to the specific needs of people with disabilities, or at the other extreme, generic, such as plug-ins to web browsers which permit interaction with common images on the web via a dialogue interface.

2 Experimental Implementation

The whole concept of communication images is implemented within the GATE project – Graphics Accessible To Everyone [8]. The server is designed as a modular component-based Java enterprise application which provides session-oriented remote services available through *SOAP* and *RESTfull* APIs. It consists of the following three modules, as shown on the UML component diagram in Fig. 1.

The **SVG Module** enables the user to upload an image and to inspect its graphical data by going through the SVG DOM tree. Either SVG or raster images can be uploaded. Raster images are automatically wrapped with initial SVG content. The suggested *Image Recognition* interface is used to extend the abilities of the module with automatic image recognition, which is very useful especially in the case of ordinary images, e.g. photos, that have no semantic data embedded so far. Image recognition algorithms must be provided by external component and connected to the *SVG Module.* Although the image recognition

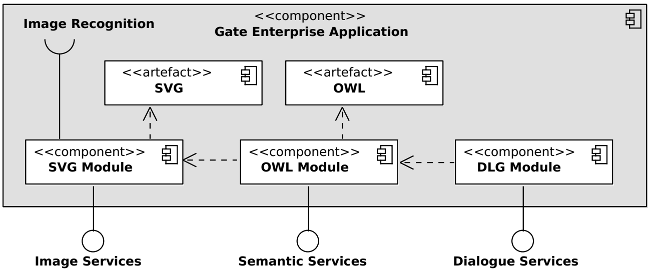

Fig. 1. Component architecture of GATE server

and auto detection techniques are still far from being able to fully describe an analyzed picture in general, specific domains, e.g. face recognition [2,13] or similarity search algorithms in large image collections [6,1] are applicable for the initial image content retrieval and can be integrated to our system.

The **OWL Module** provides semantic-related services. This module cooperates with the SVG module to get annotation data stored directly in the image and to associate them with the available knowledge database. Both the annotation data and the knowledge database are described by OWL. The implemented services cover the ontology management, low-level traversal of the OWL DOM tree, ontology reasoning and information filtering.

The **DLG Module** is responsible for the dialogue subsystem, i.e. parsing and understanding questions in natural language and then composing answers. This module cooperates closely with the OWL Module to analyze the meaning of words. At present, a simplified version of this module is implemented. This version supports questions in *What-Where Language*, WWL [8], having the format "where is what", "what is where" or "what some object is". Moreover, the engine can be configured for domain-specific utterances enabling the user to ask picture-specific questions, e.g. questions on family relationships.

Several ontologies have been proposed and integrated to the GATE project so far. *Graphical ontology* [12], for instance, prescribes important global visual characteristics of the objects such as their unusual size, dominant color or significant shape, enabling the dialogue module to express significant or unusual visual features of the objects depicted. The graphical ontology also supports the description of location and mutual position of objects (e.g. "at the upper left corner", "on the right side of another object", etc.).

Graphical ontology represents upper ontology (also known as a top-level ontology or foundation ontology [16]) describing general concepts that are the same across different knowledge domains. However, it does not handle the non-visual information required to understand the meaning of the individual depicted objects. This kind of information is supported by the domain ontologies defining vocabulary and knowledge base for concrete knowledge domain. We have developed several domain ontologies. For example, *Family* ontology can be used to classify people by their family relationships as well as infer implicit relationships. *Sights* ontology provides vocabulary and background knowledge to describe interests, historical buildings, monuments.

3 Communicative Images in Assistive Technologies

A dialogue with the image held in natural language makes the graphical data accessible especially to visually impaired people. The users are not limited to a simple summary of the image's content. Since the data is structured and related to different parts, objects and aspects of the image, a complex dialogue can be undertaken, ultimately leading to a more natural and fulfilling experience of the users.

The nature of spoken dialogue is also suitable for improving the accessibility for other groups of users with special needs. Elderly people and people with lower technological literacy would benefit from the ease of access to the information in and about the image provided by the dialogue system. The desired information does not have to be obtained manually, which might prove to be difficult for them, but on the basis of a simple request in natural language. This is also useful to motor-impaired people, people with dyslexia and some other cognitive disorders.

Moreover, a cloud-based solution enables the integration of communicative images into social media sites. These technologies support easy information sharing and on-line user collaboration, which helps to manage the knowledge in a decentralized way. The activity of one user publishing some historical facts about a monument, for example, can be utilized by other users to improve their exploration of a photo downloaded from the web. For communicative images, this kind of crowdsourcing presents an efficient way of building up knowledge bases with a long term perspective and making graphical data more accessible to everyone.

Another aspect of a collaborative structure is the social element – communication over images of mutual interest offers sharing knowledge, avocations, contacts, relations and leads to increase in a general social cohesiveness rather than the typical social inadequacy and exclusion.

These functions utilizing the ontology based information not only fulfill the user's need for information, but they also help them to exercise their memory, perception and other cognitive functions. The users with neurological or cognitive dysfunction can browse family pictures while being reminded of the age and names of the people in the photos, their birthdays, names of family pets, the time and occasion the picture was taken. Therefore, apart from the advantages concerning the access to information, communicative images can play inconsiderable role in the development of the psycho-social domain.

4 Evaluation

To evaluate the usability of the concept of communicative images, we prepared a simple experiment, where the users was aimed at exploring a given photo

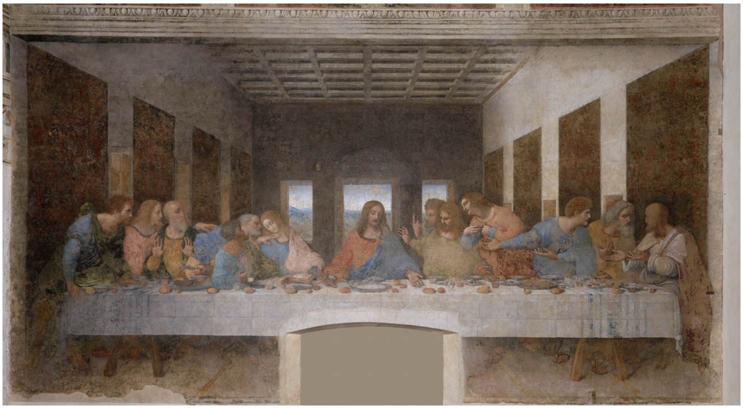

Fig. 2. Annotated painting: Last Supper by Leonardo da Vinci

by means of What-Where Language. We chose the Last Supper by Leonardo da Vinci as the reference picture, as shown in Fig. 2. This picture was precisely annotated by hand and then inspected by users via the GATE system. Da Vinci's Last Supper is one of the world's most famous paintings capturing the final meal that Jesus shared with his Apostles in Jerusalem before his crucifixion. The picture consists of several dominant objects: Jesus, 12 Apostles and a table. These objects was linked with thorough semantic data. The annotation data include historical facts about Jesus and his Apostles, their position in the picture, mutual position of the figures and the table, dominant colors of clothes, activities of figures (e.g. sitting, talking, etc.) as well as expression on their faces (e.g. curious, angry, etc.). The annotation of important figures, namely Jesus, Judas and Peter, went into more details. Their annotation was supplemented with the position of hands, for instance. Besides the annotation of picture content, the facts about the painting itself is included, e.g. when the painting was completed, its size, etc. The experiment was attended by several students of the Masaryk University with different IT skills. All the participants was sighted, therefore we hide the picture from them. Before the testing, they have been informed about the principles of communicative images. To help us understand the nature of the interaction, they have been requested to speak aloud as they communicate with the image, which enables us to record their intents, expectations or things that surprise them. After the testing, the participants filled in a questionnaire which provided us with quantitative as well as subjective evaluation.

During the experiments we used two modes, with and without guidance. Without guidance, the participants had no idea what is depicted. They communicated with the picture in any way they liked without any support. On the contrary, users with guidance was instructed what pieces of information are present and how they are structured.

Following fragment of dialogue was recorded during the free communication without guidance. This example represents a typical questions, answers and limitations in the communication (our comments are surrounded by square brackets):
***User:** What is in the picture?*
System: In the picture, there are Andrew, Table, Philip, Peter, ... [A long list of all objects depicted follows. The user would prefer a brief description instead.]
***User:** What is in the top left corner?*
System: In the top left corner there is no object.
***User:** What is in the middle?*
System: In the middle there are James Major, Jesus, Thomas, Peter, Philip, John and Judas. [User asked for various locations one by one. Then he was wondering about the exact order of figures. Some questions failed at this moment.]
***User:** What is on the left side of Jesus?*
System: On the left side of the Jesus are Andrew, Peter, Bartholomew, Judas, Right Hand of Judas, James Minor and John.
***User:** Is Peter on the left side of Andrew?*
System: The position ”left side of andrew” is not valid for searching for object Peter. Please, re-type your query.
***User:** Is Peter in the left of Andrew?*
System: The position ”left of andrew” is not valid for searching for object Peter. Please, re-type your query. [Several attempts continued.]
***User:** Describe Table.*
System: The table around which are Jesus and all the Apostles sitting. [This is misunderstanding. In fact, the user was interested in objects on the table, not description of the table.]
***User:** Describe Jesus.*
System: Jesus (72 BC to 3033 AD), also referred to as Jesus of Nazareth, is the central figure of Christianity, whom the teachings of most Christian denominations hold to be the Son of God. Christianity regards Jesus as the awaited Messiah of the Old Testament and refers to him as Jesus Christ, a name that is also used in non-Christian contexts. He is also a major figure in Islam. In the painting, Jesus appears to be calm, tranquil. [Another misunderstanding. The user would prefer to put emphasis to the last sentence, i.e. it would be better to distinguish between “Describe Jesus” and “Who is Jesus”.]

In what follows, there is a transcription of guided dialogue, i.e. the user knew what kind of information is available:
***User:** Describe the picture.*
System: Object cannot be recognized. Please, re-type your query.
***User:** Detailed information about picture.*
System: This photo presents the painting called ”The Last Supper”, which portrays the reaction given by each apostle when Jesus said one of them would betray him.
***User:** Who painted the picture?*
System: This painting was created by Leonardo da Vinci.
***User:** What people are in the picture?*

System: Sorry, I don't understand your query. Please, enter another query.
User: *Who is in the picture?*
System: In the picture, there are Andrew, Philip, Peter, Bartholomew, James Major, Judas, Thomas, Thaddeus, Simon, Matthew, James Minor, John and Jesus.
User: *Where is Judas?*
System: Judas is in the middle.

Subjective impression of all the participants was very positive. All the people were able to recognize the painting, reconstruct the layout in their mind and imagine the scene. However, we have to point out that the sighted people participating in the experiment were probably familiar with this popular painting. For blind users this task could be more difficult.

Users usually assessed the interaction with the image as funny, quite easy and natural. On the contrary, effectiveness of the dialogue based interaction varied from "very effective" to "not very effective". Users frequently criticized weak misunderstanding solving, the lack of hints and missing support of general questions that do not belong to the scope of WWL.

5 Conclusions and Future Work

In this paper, we have outlined basic principles of communicative images as well as the general architecture of the system. The aim of the performed experiment was to make sure that the concept is viable, implementable and useful. The implementation is still very simplified and the concept of communicative images has many open problems. For instance, there is the gap between semantic models and dialogue strategies. At the moment, we have to carefully prepare and fine-tune the dialogue subsystem for each concrete domain ontology by hand instead of generating dialogue strategies automatically from the internal structure of provided ontology. Also continual enhancement and enlargement of knowledge base as well as automatic learning from dialogues pose a big challenge.

The preliminary results show that this approach promises valuable utilization in many application domains like e-learning, smart management of large photo collections or assistive technologies. However, more experiments especially with visually impaired people have to be performed to verify feasibility of communication images.

References

1. Abbasi, R., Chernov, S., Nejdl, W., Paiu, R., Staab, S.: Exploiting flickr tags and groups for finding landmark photos. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 654–661. Springer, Heidelberg (2009)
2. Bartlett, M., Movellan, J.R., Sejnowski, T.: Face recognition by independent component analysis. IEEE Trans. on Neural Networks 13(6), 1450–1464 (2002)

3. Dahlström, E., et al.: Scalable vector graphics (svg) 1.1, 2nd edn. (2011), http://www.w3.org/TR/SVG/
4. Faro, A., Giordano, D., Spampinato, C.: Combining literature text mining with microarray data: advances for system biology modeling. Briefings in Bioinformatics 13(1), 61–82 (2012), http://bib.oxfordjournals.org/content/13/1/61.abstract
5. Guarino, N.: Formal ontology and information systems, pp. 3–15. IOS Press (1998)
6. Jaffe, A., Naaman, M., Tassa, T., Davis, M.: Generating summaries and visualization for large collections of geo-referenced photographs. In: Proc. of ACM Int. Workshop on Multimedia Information Retrieval, New York, USA, pp. 89–98 (2006)
7. Kopecek, I., Ošlejšek, R., Plhák, J.: Dialogue management in communicative images. In: Text, Speech and Dialogue - Students' section, Proceedings Addendum, pp. 9–13. University of West Bohemia in Pilsen in Pilsen, Publ. House, Pilsen (2011)
8. Kopeček, I., Ošlejšek, R.: GATE to accessibility of computer graphics. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 295–302. Springer, Heidelberg (2008)
9. Kopecek, I., Oslejsek, R.: Communicative images. In: Dickmann, L., Volkmann, G., Malaka, R., Boll, S., Krüger, A., Olivier, P. (eds.) SG 2011. LNCS, vol. 6815, pp. 163–173. Springer, Heidelberg (2011)
10. Lacy, L.W.: OWL: representing information using the Web Ontology Language. Trafford Publishing, Victoria BC, Canada (2005), http://www.worldcat.org/search?qt&=1412034485
11. Linton, F., Joy, D., Peter Schaefer, H.: Building user and expert models by long-term observation of application usage. In: Proceedings of the Seventh International Conference on User Modeling, pp. 129–138. Springer (1999)
12. Ošlejšek, R.: Annotation of pictures by means of graphical ontologies. In: Proc. Int. Conf. on Internet Computing, ICOMP, pp. 296–300 (2009)
13. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. In: Proceedings of the 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1996, pp. 203–208 (June 1996)
14. Satria, H., Priya, R.S., Ismail, L.H., Supriyanto, E.: Building and reusing medical ontology for tropical diseases management. International Journal of Education and Information Technologies 6, 52–61 (2012)
15. Shah, N., Jonquet, C., Chiang, A., Butte, A., Chen, R., Musen, M.: Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics 10(suppl. 2), 1–10 (2009), http://dx.doi.org/10.1186/1471-2105-10-S2-S1
16. Staab, S., Studer, R.: Handbook on Ontologies, 2nd edn. Springer Publishing Company, Incorporated (2009)
17. Wang, Y., Cena, F., Carmagnola, F., Cortassa, O., Gena, C., Stash, N., Aroyo, L.M.: RSS-based interoperability for user adaptive systems. In: Nejdl, W., Kay, J., Pu, P., Herder, E. (eds.) AH 2008. LNCS, vol. 5149, pp. 353–356. Springer, Heidelberg (2008), http://dblp.uni-trier.de/db/conf/ah/ah2008.html#WangCCCGSA08