Vind(x): using the user through cooperative annotation

Louis Vuurpijl
NICI, University of Nijmegen, The Netherlands
E-mail: vuurpijl@nici.kun.nl

Lambert Schomaker
RUG, University of Groningen, The Netherlands
E-mail: schomaker@ai.rug.nl

Egon van den Broek
NICI, The Netherlands
E-mail: e.vandenbroek@nici.kun.nl

Abstract

In this paper, the image retrieval system Vind(x) is described. The architecture of the system and first user-experiences are reported. Using Vind(x), users on the Internet may cooperatively annotate objects in paintings by use of the pen or mouse. The collected data can be searched through query-by-drawing techniques, but can also serve as an (ever-growing) training and benchmark set for the development of automated image retrieval systems of the future. Several other examples of cooperative annotation are presented in order to underline the importance of this concept for the design of pattern recognition systems and the labeling of large quantities of scanned documents or online data.

1 Introduction

Image retrieval has become an increasingly popular research theme during the last years. This is no surprise, given the vast amounts of electronically available image archives and the rapidly increasing accessibility of large digital image collections (e.g., the Internet, press photo databases, museum collections, and so forth). In the cultural domain, museums like the Dutch Rijksmuseum at `www.rijksmuseum.nl` or the Hermitage at `www.hermitage.ru`, are extending their reach by making part of their collection publicly available via the Internet. Many of such initiatives are now undertaken by libraries, museums and governmental institutes with the goal to preserve our cultural heritage. Though this opens up new possibilities for sharing and distributing image data, it also creates the need for information systems for indexing, browsing, and retrieval of visual information.

Unfortunately, the image retrieval systems currently available are far from mature, as they still yield unacceptable retrieval results, are restricted in the domain that is covered, lack a suitable user-interface and are mainly technology-driven, requiring a lot of domain knowledge from the user about how to install features that will fulfill their information need [6, 7]. Therefore, research on this topic has seen a shift from computer vision and pattern recognition to other disciplines such as cognitive science and psychology. For example, Rui *et al* [6] and Jörgensen [2] emphasize that it is paramount to consider "the human in the loop". Using knowledge about the user will provide insight in how the user-interface must be designed, how retrieval results may be presented, and it will categorize the typical information needs that are shared by the general public.

With that in mind, a large Dutch project called ToKeN2000 (see `www.token2000.nl`) has been initiated as an interdisciplinary research programme that combines seven research institutes with an affiliation in computer science and cognitive science. Major theme is the improvement of accessibility and retrieval of knowledge, focusing on fundamental problems of the interaction between a human user and an information retrieval system. As an experimentation platform, the digital collection of the Dutch Rijksmuseum is used, comprising a large database of 60000 paintings. In the frame work of ToKeN2000, we have developed a web site `http://kepler.cogsci.kun.nl/vindx` through which we want to assess and integrate the technological and usability aspects required for the design and implementation of successful image retrieval systems.

The single most important consideration for the design of Vind(x)[1] was based on the observation that the majority

[1] The word *vind* is the Dutch equivalent for *find*.

of users seeking visual information, is looking for image material containing a specific *object*. To pursue this question ("what information are you looking for?"), an inquiry was held among users on the Internet. About 72% of the respondents considered images with, e.g,. a dog, a human, or a house, as more interesting than images depicting a certain scene [7]. The participants cooperatively contributed to this outcome and we will argue in this paper that this approach, that *uses the user*, can be applied to a wide variety of tasks that are of interest for the pattern recognition community.

Unfortunately, the automated unconstrained recognition of objects in image material remains a largely unsolved problem. Although many papers in the literature describe a (partial) solution, these all comprise a small number of objects for a very limited domain. Furthermore, in order to build a proper object model, training data is required. This is the general problem of statistical pattern recognition, where recognition performance is directly related to the availability and quality of the training data. Note that, whereas machines are still incapable to do so, humans can very well perform the required distinction between an object and its background. So, why not use the user for gathering supervised data? In this paper it is described how, through the concept of cooperative annotation, "the presence of human perceptual abilities" [7] is exploited to generate a dataset comprising outlines of objects in images together with their corresponding class labels and textual descriptions.

The rest of this paper is organized as follows. In the next section, we elaborate on the concept of cooperative annotation. A number of examples is given that support our opinion that users can effectively be used to help in the design of a system, to collect supervised data that can be used to train pattern recognition algorithms, or to investigate what constitute the salient aspects of a system as considered by its intended users. In Section 3, the architecture of Vind(x) is described and an example of cooperative annotation for image retrieval is given. Section 4 presents an ongoing experiment that investigates the use of *query-by-drawing*. In this experiment, users are requested to draw the closed contour of an object, that can be used to query the database containing already collected outlines of objects. It will be assessed whether query-by-drawing is a usable technique for image retrieval.

2 Cooperative annotation

The SETI initiative (search for extra-terrestrial intelligence) is an excellent example of the joint exploitation of available resources. Users on the Internet are making their computers available to cooperatively explore massive amounts of radio telescope data with the goal to look for alien life. Similar number crunching efforts have been reported for, e.g. computing a world record large prime number or cracking the RSE DES-II key. A cooperative effort that uses the *user* rather than mere computing resources is OpenMind [8]. On the web site `openmind.org`, it is explained that *"The Open Mind Initiative is a collaborative framework for developing "intelligent" software using the Internet. Based on the Open Source method, it supports domain experts (who provide algorithms), tool developers (who provide software infrastructure and tools) and non specialist "e-citizens" (who contribute raw data)."* This initiative, launched in 1999, now uses users from the speech recognition, handwriting recognition and other communities for its goals.

More recently, a cooperative document understanding system was described in [5]. In their system, called Edelweiss, multiple users are allowed to access and annotate the same document, thus cooperatively joining expertise to establish the task at hand. Downton *et al* describe a legacy document conversion system [1] that scans huge stacks of handwritten index cards from, e.g., manually organized museum archives. The scanned documents are processed using OCR techniques to generate an online archive. This archive is accessible through the web and users are contributing to the project by interactively validating the content as it is used.

We have used this concept in our lab since 1995 for a number of tasks related to handwriting recognition and information retrieval. Below, three of these tasks are described to further point the reader at the impact that cooperative annotation can have, in particular with respect to the labeling of scanned (offline) images or online handwriting data.

2.1 Web-based annotation of scanned images

The NICI has been involved in a comparison study of two forensic writer identification systems. In order to be able to assess the recognition performance of both systems, a suitable benchmarking data set had to be defined and collected. During four different writing conditions, a total of 250 subjects were asked to produce (i) constrained normal texts, (ii) constrained block capital texts, (iii) constrained forged handwriting and (iv) unconstrained texts. The production of constrained handwriting involved that subjects had to copy a number of pre-defined lines of text. Unconstrained handwriting was collected by showing the subjects a cartoon, which they had to describe in their own words.

The collected and scanned data had each to be examined on two issues. First, it had to be verified whether the three constrained sets were correctly copied. Second, the texts that were produced during the unconstrained condition had to be labeled.

We were able to cooperatively perform this labor inten-

sive job (quality control of 750 images, labeling of 250 images) within 20 days as follows. A dedicated web-server hosted all collected images of scanned handwriting. On each day, all five participants received 10 emails containing an URL they had to visit. Upon visiting the URL, they were presented with an image that had to be verified and/or labeled through a web-form.

2.2 Web-based description of image content

As part of a graduation project [3], we developed a web site that hosted a database of several hundred images, collected from the Internet and with content from various domains. Upon visiting the site, a random image was presented to the visitor. Through a web form, the visitor was asked to describe the content of the image in an unconstrained fashion. As it happened, the student involved was one of the developers of a very popular web site that attracted many visitors. Via this site, users were kindly requested to participate in our project and within several months, textual annotations of image material were collected from well over 22,000 participants. Note that these texts contain semantic descriptions of image material as observed by "e-citizens", comprising a valuable data set that will be explored in the near future.

2.3 Web-based labeling of handwritten words

We have developed a handwriting recognizer, called dScript, that has been on display at two Dutch museums and the IWFHR7 conference. All data written by the users of dScript was stored. The data contains thousands of Dutch city-names and is labeled by the recognizer. We are currently verifying the labels through cooperative annotation. Users from our lab can contribute by clicking an URL, after which an image of a word together with the top-ten hit list of the recognizer are displayed. The verification of the labeling is done by (i) selecting the correct word from the list, or (ii) entering the correct label via a text-entry field.

2.4 Quality control

There is an important issue about quality control associated with web-based annotations. For the first and third example, we relied on a small number of trusted users from our own department. For the collection of textual descriptions of image content, we have no way to ensure that the annotations indeed reflect the image content. However, our experience with the collected data through dScript as well as the data acquired through Vind(x) indicates that more than 95% can be marked as cooperative. Furthermore, the process of cooperative annotation itself could be used for the purpose of quality control, where each annotation produced by an untrusted user is verified by someone from our lab. And, in case we would decide to allow "e-citizens" to label handwritten words, two or more annotations of the same word could be compared automatically. If the labels do not match, a trusted person could be asked to rule which label is correct.

2.5 Basic architecture of web-based cooperative annotation

All three examples presented in this section use the same architecture, as depicted in Figure 1. A database server hosts a set of unlabeled objects, e.g. scanned documents, digital photographs, or handwritten words. This is called the *object database* and request for objects from the database are marked as solid arrows in Figure 1. Annotations collected via cooperative annotation are stored in the *annotation database* (marked with dotted arrows). Upon clicking a URL through his browser, the web server contacts the database server, requesting for a new object to be annotated. Web forms or Java applets are used to collect the annotations of an object.

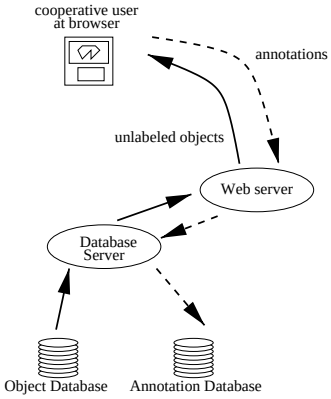

Figure 1. General architecture of web-based cooperative annotation.

3 The Vind(x) system: pen-based annotation

Vind(x) is an image retrieval system that:

- allows browsing through a subset of the digital collection of the Rijksmuseum,
- introduces a novel way of information presentation,
- provides an interface for cooperative annotation of paintings, and
- implements the concepts of query-by-drawing, query-by-example and text-based querying [6, 7].

The domain that is covered by Vind(x) contains 17th century paintings. Vind(x) comprises an image browser via which the user can step through the collection. For each painting, an external link to the web site of the Rijksmuseum is added and the user is presented with information about how many people annotated the painting. Using the concept of mouse-over events, the outlines and textual descriptions from annotated objects can be revealed in attractive way that is rated as considerably informative by the users. Figure 2 depicts part of the user-interface, after a user has outlined a bird and has entered the required descriptions.

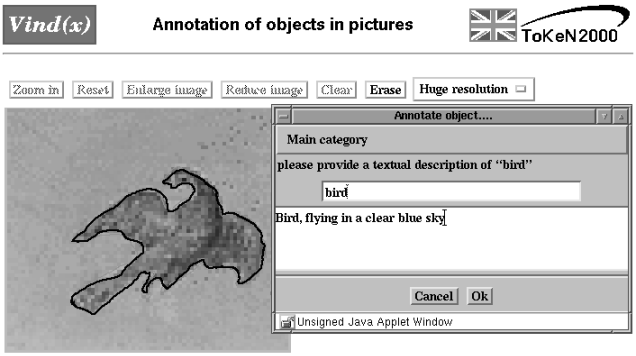

Figure 2. Annotating a flying bird. The user has first produced an outline enclosing the object. Subsequently, the user is asked to describe the object with one word and with a more elaborate textual description.

For the annotation of objects in paintings, Vind(x) provides a Java interface. Using the applet, a user can zoom in on interesting parts of the painting and start drawing a closed outline surrounding the object of his/her interest. When finished drawing, the user is requested to provide (i) the object class, e.g., person, plant, animal, (ii) a one-word description and (iii) an unconstrained textual description of the object. The same architecture as depicted in Figure 2 is used for the cooperative annotation process of Vind(x).

3.1 Querying the object database

For querying objects from the database, a separate process is running that is able to interpret and respond to requests from browsers on the internet. This process, called the *query server*, is able to transmit requests to a number of specialized query engines, or agents, that handle one of the specific querying paradigms of Vind(x). This architecture, depicted in Figure 3, builds on the agent architecture described in [9]. In that paper, we introduced a framework for combining the expertise of several experts in a distributed system.

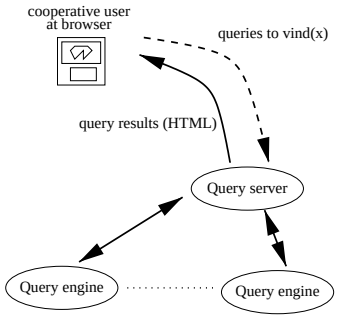

Figure 3. General architecture of the Vind(x) querying system.

Vind(x) implements several querying paradigms. Based on the annotated textual descriptions, plain text-based or categorical search may be used to retrieve images containing a specified object. Another possibility to query the database of outlines is to present a set of thumbnail images of annotated objects, from which an image that is similar to the user's information need can be chosen. Subsequently, the system retrieves images that match the example. This technique, called query-by-example, is most often used by image retrieval systems. It implicitly uses human perception as a selection mechanism to navigate through the document search space.

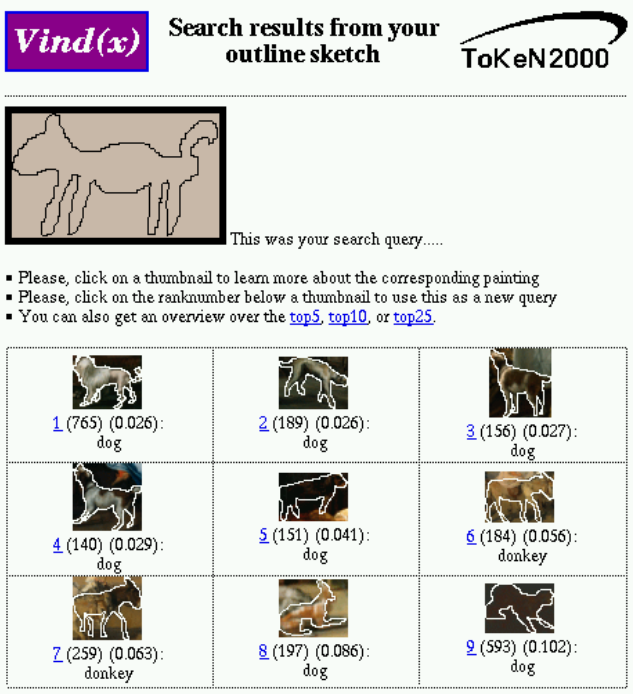

Figure 4. Example of query-by-drawing

A third querying paradigm, query-by-drawing, is depicted in Figure 4. Using the pointer, a user is free to draw a closed outline, which is matched to the outlines stored in the database using outline matching [7].

3.2 First user experiences

Vind(x) has been online since the beginning of 2001. It was extensively tested by people from our lab, and thousands of visitors were recorded to have browsed the system. In total, 3000 outlines have been collected. We have explored the concept of annotating images with a pointer since several years [7] and concluded that users are able to produce usable outlines. Considering the outlines from the Vind(x) database, this conclusion is further justified. From usability studies through observation, it appears that users like the way in which retrieval results are presented. Even if the system makes mistakes, users can understand and accept why this is the case, as apparently shape-based matching yields results that are visually perceptive to the human user [4]. As an example of this effect, consider Figure 4. The user was looking for dogs, but somehow two donkeys from the database matched his query well. As both object classes are visually similar, users are less frustrated than in the case of miss-matches when other feature schemes, such as color distributions are used. We have obtained similar experiences with usability studies of handwriting recognizers, where users accept recognition errors when it is shown or explained how the system reaches a wrong decision.

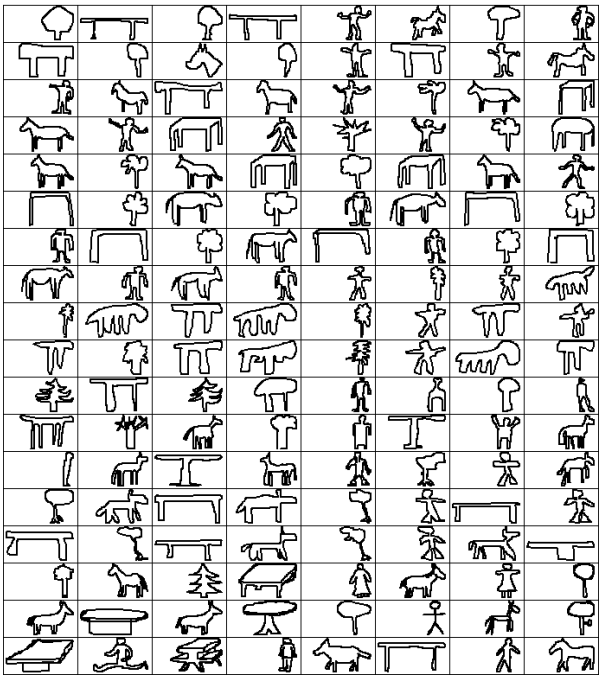

Figure 5. Example outlines drawn by heart.

Note that the outlines contained in the database have been produced by tracing the contours of existing objects in images. This means that the resulting shapes are determined by the photographer or painter of the image, as explained in [7]. In order to further assess the concept of query-by-drawing, we are currently conducting an experiment that uses outlines drawn by heart.

4 Limitations of query-by-drawing

The two questions that are addressed here are: are users capable of producing closed outlines? and can the produced outlines be used as a query to retrieve specific objects from our database?. Via the web page `http://loop.cogsci.kun.nl/egon/exp1/`, human subjects were requested to produce 5 instances of 4 different objects (a person, a horse, a table and a tree). For each object, a closed outline had to be drawn. The preliminary results presented here, are based on 520 samples drawn by 26 students, as shown in Figure 5. The subjects had to complete an evaluation form that would yield feedback on whether they encountered any difficulties. Five students mentioned that they were very limited by the requirements that the outline had to be closed and that no pen-lifts were allowed. However, all participants reported that drawing with the mouse was feasible, in particular those with a longer experience in computer usage. From Figure 5 it can be observed that certain users exploit a far more developed artistic skill than others.

To asses the second question, two pattern recognition tasks were performed using the matching algorithms described in [7]. The first used each of the 520 samples as a query from the sample database. The second used each sample as a query to the entire database comprising the original 3000 samples. The original database contains 87 humans, 1 table, 2 donkeys and 8 trees. Below, the classification results are shown, for the top-1, top-5 and top-10 retrieval lists.

Table 1. Retrieval results for querying a horse, table, tree or human.

	520 samples			3000+520 samples		
	top-1	top-5	top-10	top-1	top-5	top-10
tree	96.7	92.0	86.7	91.7	70.3	53.5
human	93.3	87.3	84.5	95.0	89.3	86.0
table	91.7	82.3	72.7	91.7	78.7	68.7
horse	98.3	95.0	89.5	91.7	88.7	83.2

Each cell from the table indicates the percentage of retrieved cases in a list. Query objects were excluded from the retrieval list. For the first experiment, table 1 shows that in

96.7% of the cases where a tree was used as query, the first item retrieved was a tree. The second column shows that on average 4.6 trees were in the top-5 list and the third column indicates that 8.7 trees can be found in the top-10 list.

The vast majority of hits represents objects that match the intended information need. The results from retrieving objects from the larger database are still rather good. In more than 92% of the cases where a specific object was searched, an object belonging to the same class was retrieved. Again, query objects where excluded from the lists. The worst case occurs when retrieving trees, resulting in little over five relevant hits in a top-10 hitlist.

However, using the concept of query-by-example, these hits could be used by the user in an intuitive manner to zoom in on his information need. By clicking on one of the retrieved objects, the user would be able to specify which object matches his request the best. Furthermore, if it would be recorded that a significant amount of users use tree-shaped outlines as a query, our matching routines could be re-designed to meet demands from actual user usage.

5 Conclusions and future directions

Cooperative annotation was identified as an important paradigm that *uses the user* for the collection of training data in domains where the machine-based recognition of, e.g., objects in images and online or offline handwriting, is largely unsolved. We have presented a general architecture of web-based annotation systems and demonstrated several scenarios where this concept has proven to be very successful. People are willing to participate in web-based experiments, as was shown by the example where more than 22,000 users participated in the acquisition of textual descriptions of image content. The automated quality control of such a collection remains a challenging issue that we will pursue in the future, though it was indicated that the cooperative verification by a number of trusted users may yield a first step towards this goal.

It was argued that current image retrieval systems are mainly technology driven and that incorporating knowledge about the user is vital for the successful application of novel retrieval techniques like query-by-drawing. The architecture and first user experiences of Vind(x), a web site that uses cooperative annotation for indexing the digital collection of the Dutch Rijksmuseum were presented. It was shown that users are willing to cooperate by annotating objects in images. Usability studies have indicated that the way in which visually perceptive retrieval and information presentation techniques are implemented in Vind(x), were particularly rated as appealing by the users.

Within the frame work of ToKeN2000, we will pursue the challenge of combining outline-based features with "traditional" image features such as color and texture. The data collected through Vind(x) will provide a valuable source of information that will certainly help to further design and test techniques for automatically detecting objects in images.

References

[1] A. Downton, A. Tams, G. Wells, A. Holmes, and S. Lucas. Constructing web-based legacy index card archives – architectural design issues and initial data acquisition. In *Sixth International Conference on Document Analysis and Recognition*, pages 854–864. IEEE, September 2001.

[2] C. Jörgensen. Access to pictorial material: A review of current research and future prospects. *Journal of Human Computer Studies*, 44(6):875–920, 1999.

[3] M. Koenen. Image retrieval through natural language queries. Master's thesis, Nijmegen Institute for Cognition and Information, 2002. In preparation.

[4] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[5] N. Roussel, O. Hitz, and R. Ingold. Web-based cooperative document understanding. In *Sixth international conference on document analysis and recognition*, pages 368–373. IEEE, September 2001.

[6] Y. Rui, T. S. Huang, and S. Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, 1999.

[7] L. Schomaker, L. Vuurpijl, and E. de Leau. New use for the pen: outline-based image queries. In *Fifth International Conference on Document Analysis and Recognition*, pages 293–296. IEEE, September 1999.

[8] D. Stork. Character and document research in the open mind initiative. In *Fifth International Conference on Document Analysis and Recognition*, pages 1–12. IEEE Computer Society, September 1999.

[9] L. Vuurpijl and L. Schomaker. Multiple-agent architectures for the classification of handwritten text. In *IWFHR6, International Workshop on Frontiers of Handwriting Recognition*, pages 335–346, August 1998.