An Intelligent Web Crawler

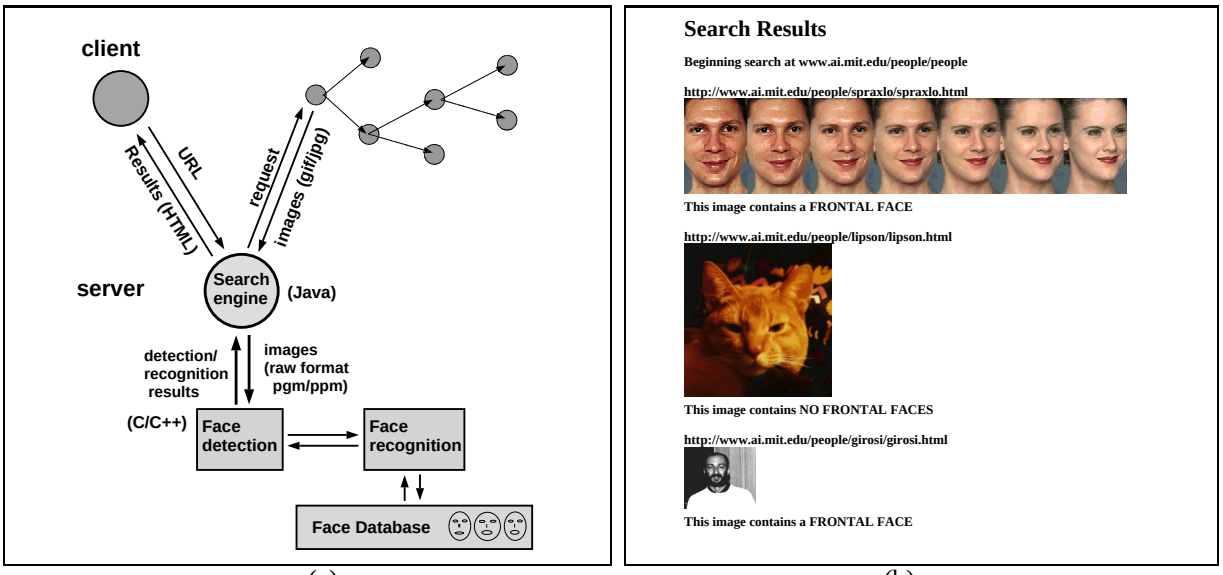
Michael Oren and Tzu-Mainn Chen

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: To develop an intelligent search engine for the World-Wide-Web which is based on the analysis of the image content of various sites.

Motivation: The problem of information search on the World-Wide-Web has become of critical importance with its enormous growth in the recent years. Common search engines and site indices (e.g., Alta-Vista, Yahoo, Excite and others) which have been developed to address this problem are based mainly on textual analysis of the site content, (e.g., word statistics,) or on categorical classification (e.g., business, sports, magazines, etc.) In contrast to the text-based approach, we have developed a search engine based on visual content of the various sites. The search engine based on recently developed computer vision techniques enables the user to locate sites containing images of certain types (e.g., photographs of people or cars,) or even to search for a particular person. By incorporating computer vision techniques into a Web search engine, the user's ability to search information is extend beyond the the capability of text-only analysis.



(a) (b)

Figure 1: (a) System Architecture (b) An example of a search result. The results page contains the images and the corresponding URL addresses. The URL addresses are also Web links to the original sites.

System Description: The architecture of the search engine is based on the Web-crawler scheme and is shown in Figure 1(a). The client (the user) sends to the server (the search engine) a URL address that is the starting point of the search. The Web crawler visits the given site, locates all the images, and retrieves them to the server. The server converts the images to a standard format and sends them to the image analysis module which can employ various computer vision algorithms. Currently, the system uses a face detection algorithm, [1][2], which detects whether or not an image contain frontal faces, even if the background is complex and cluttered. Once the search engine analyzes the images of the current site, it identifies all the links to other sites and continues the search in a breadth-first manner

until a predefined number of sites were visited.

The server sends back to the client's Web browser the results of the search in HTML format which include the images of interest and the addresses of the corresponding Web pages. The result page contains links to the original sites which provide the user with easy navigation and browsing capabilities. A typical result of of the search engine with the face detection module is given in Figure 1(b).

Currently, a face recognition module, [3], is added to the system which will enable it, not only to search for images with faces, but also to locate a particular person, by comparing the located faces with the user's database. Every user will be able to build her own database of faces simply by selecting a set of face images from the Web.

System Implementation: The Web crawler engine (the server) and the image analysis processor are designed as two independent modules with a minimal interface. The server which handles the interface with the user and the Web is implemented using the Java programming language. The image analysis module is implemented using C/C++ since it is computationally intensive. The hybrid architecture of Java and C/C++ allows the system to be compact and modular without compromising performance. Each one of the modules can be modified independently of the other and different image analysis techniques can be added easily. The search engine was installed as CGI script which is accessible to every user on the Web.

Impact: The intelligent Web crawler enables users to search the Web not only based on textual information and subject classification but also based on its visual content. The Web crawler, as a front-end to vision applications, will be able to provide a variety of features and will help to transform the Web into huge and easily-accessible database of images. The different users will be able to easily retrieve images of interests. For example, one can search the Web for photographs of friends, an architect may utilize it to search for different building designs, and a racing fan will probably use it to collect images of cars. By augmenting Web search engines with computer vision techniques, the ability of users to manage and filter data on the Internet is significantly enhanced and new applications that combine text and visual information can be developed.

Future Work: Future extensions include adding new object detection algorithms, [4], and detecting a larger variety of object classes such as pedestrians, pets and cars. We also plan to add a trainable scene classification algorithm, [5], which will enable the user to define a qualitative scene concept, e.g. snow-capped mountains, by a set of examples. Once the concept is defined the Web crawler can be used to search for new images describing the same type of scene.

Other Collaborators: Jason Miller and Charles Lee.

References

[1] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *MIT AI Lab-Memo*, No. 1521, 1994.

[2] Edgar Osuna, Robert Freund, and Federico Girosi. Support vector machines: Training Support Vector Machines: an Application to Face Detection. *Proceedings of CVPR-97*, June 1997. .

[3] R. Romano, D. Beymer, and T. Poggio. Face verification for real-time applications. *Proceedings of the 1996 Image Understanding Workshop*, February 1996.

[4] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. *Proceedings of CVPR-97*, June 1997.

[5] P. Lipson, E. Grimson, and P. Sinha. Configuration based scene classification and image indexing. *Proceedings of CVPR-97*, June 1997.