

An Overview of the WeOCR System and a Survey of its Use

Hideaki Goto

Information Synergy Center, Tohoku University, Sendai, 980–8578 Japan.

<http://www.sc.isc.tohoku.ac.jp/~hgot/>

email: hgot@isc.tohoku.ac.jp

Abstract

The collaboration of the network and optical character recognition (OCR) is expected to be useful for extending the applications of OCR, and it has a potential to open up new vistas in future services using OCR. We presented the basic ideas of the synergetic OCR system in 2004, and developed a platform called WeOCR to realize the web-based OCR services. The number of accesses has been increasing gradually since the interoperable, synergetic web-based OCR system was put into service in 2005. In this paper, we describe an overview of the current WeOCR system and provide a survey of its use. The unique survey may convey some useful hints for future development of network-oriented OCRs, not limited to the WeOCR, and potential applications of such systems.

Keywords: web-based OCR, WeOCR, OCRGrid, character recognition, document analysis

1 Introduction

The collaboration of the network and optical character recognition (OCR) is expected to be useful for making lightweight applications not only for desktop systems but also for small gadgets and autonomous robots.

For example, some mobile phones with character recognition capabilities are available in the market today [1]. A mobile phone with a real-time English–Japanese translator was proposed by a Japanese company. Although it has a simple, built-in OCR, it requires a large language dictionary provided by an external server on the Internet. Another example is the vision system of autonomous robots. We have been working on an autonomous robot system with text detection and character recognition capabilities [2]. Due to the limitation of the hardware, it is difficult to put a large character-set data and a dictionary for language processing into small gadgets or robots. In addition, we cannot use a sophisticated character recognition method, because the processor power is very limited. Thus, use of network can be beneficial to various applications.

The network-based architecture has some advantages from the researchers' and developers' points of view as well. Recent OCR systems are becoming more and more complicated, and the development requires expertise in various fields of researches. Building and studying a complete sys-

tem has become very difficult for a researcher or a small group of people. A possible solution would be to share software components as Free Software among researchers and developers. However, many researchers think the programs are a kind of intellectual property and are not willing to release the source codes even after the related papers have been published. An alternative solution is to use Web Application Servers. Since the programs run on the server side, people can provide computing (pattern recognition) power to others without having their source codes or executables open. A prototype system for a text locating competition was proposed [3].

In 2004, we presented a new concept of network-oriented OCR systems called Synergetic OCR [4]. The basic ideas are totally different from those of some experimental websites with OCR capability that existed at that time for demonstration purposes. We have designed the Synergetic OCR to make a lot of OCR engines work cooperatively over the network worldwide to gain some synergetic effects such as performance improvement of OCR, to realize multilingual, high-function, sophisticated OCR systems, and to provide ubiquitous OCR services. We have revised the Synergetic OCR later and proposed a new conceptual platform called OCRGrid in 2006 [5].

Since year 2004, we have been designing a web-based OCR system, called WeOCR, as a simplified instance of the OCRGrid. The WeOCR system

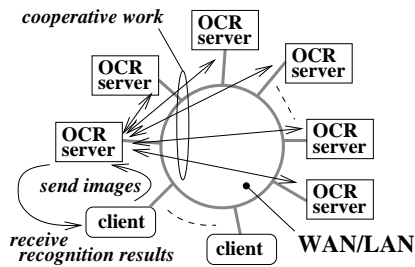


Figure 1: OCRGrid platform.

was put into service in 2005. One of the biggest concerns of a web-based OCR is about the privacy of documents. A simple question arises — Do people really want to use a system like that? To answer this question, we probably need to investigate the usage of the system.

In this paper, we describe an overview of the current WeOCR system, present the usage statistics, and provide a survey of its use to obtain some useful hints for future development of the WeOCR and potential applications of OCR.

2 Overview of the WeOCR System

2.1 Basic Concept

The OCRGrid is a conceptual platform, while the WeOCR is an implementation of the OCRGrid based on HTTP (HyperText Transfer Protocol).

The basic concept of the OCRGrid is rather simple. A lot of OCR servers are deployed on a network as shown in Figure 1. The OCR servers work either independently or cooperatively communicating with each other. A client connects to one of the OCR servers and sends text images to it. The servers recognize the images and send the recognition results (i.e. character codes, etc.) back to the client.

Although the OCRGrid is expected to be used worldwide over the Internet, we may use any networks such as wireless networks, corporate local area networks (LANs), virtual private networks (VPNs), interconnection networks in parallel servers, and even interprocess communication channels.

Some potential applications of the platform have been discussed in [4, 5].

Unlike some commercial OCR systems equipped with web interfaces, we are expecting the WeOCR servers will be supported by the communities of researchers, developers, and individuals as well as application service providers.

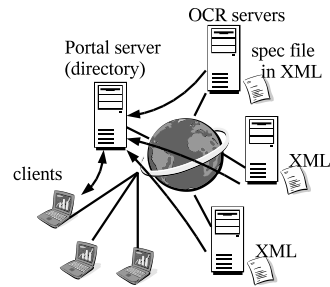


Figure 2: Directory-based server search system.

2.2 WeOCR Toolkit

Making a web-based OCR from scratch as an application server requires a lot of expertise about network programming and network security. Since many OCR developers and researchers are not so familiar with network programming, we have developed a toolkit¹ to help those people build secure web-based OCRs easily. The toolkit consists of some interface programs and filter programs that work together with the web server.

The document image is sent from the client computer to the WeOCR toolkit via the Apache web server program. The programs in the toolkit check the image size, examine the file integrity, uncompress the file if necessary, convert the image file into a common file format, and invoke the OCR software. The toolkit includes some programs used for protecting the web server from malicious attacks or from the defects of the OCR programs.

The recognition results are converted into HTML data and sent back to the client.

To enable the end users to search for appropriate OCR engines easily, we have developed a directory-based server search system. Figure 2 depicts the overview.

A specification file written by the server administrator is attached to each OCR engine. The specification file is written in XML (eXtensible Markup Language) so that automated data handling becomes easy. The file describes the specifications of the server, including the location (URL) of the server, the location of the application interface program (CGI), the name of the OCR engine used, supported languages, supported document types, etc. The portal server has a robot program for collecting the specification data automatically and periodically from the OCR servers. The robot analyzes each specification file in XML and update the database entries. A simple search program picks up the OCR servers that match the client's needs from the database and shows the search results.

The toolkit and the server search system may be easily adapted to other online applications related

¹The toolkit is available at <http://weocr.ocrgrid.org/>

Table 1: WeOCR servers (as of Aug. 2007)

Server (location)	Sub-ID	Engine	Deployed	Description
ocr1.sc (Japan)	ocrad	GNU Ocrad	Feb 2005	Western European languages
	gocr	GOOCR	Nov 2005	Western European languages
appsv.ocrgrid (Japan)	ocrad	GNU Ocrad	Jun 2006	Western European languages
	hocr	HOCR	Jun 2006	Hebrew OCR
	tesseract	Tesseract OCR	Aug 2006	English OCR
	scene	Tesseract OCR + private preprocessor	Apr 2007	Scene text recognition for English (experimental)
ocr.ekitap (Turkey)	ocrad	GNU Ocrad + private postprocessor	Aug 2006	Turkish OCR
asv.aso (Japan)	ocrad	GNU Ocrad	Aug 2006	Western European languages
	tesseract	Tesseract OCR	Aug 2006	English OCR

to computer vision and pattern recognition.

2.3 Available Servers

Table 1 shows the WeOCR servers available on the Internet as of Aug. 2007.

All the servers are constructed based on Open Source OCR software so far. Although these OCR engines work very well with some limited kinds of documents, the average performance is basically inferior to that of the commercial products. Thanks to the Tesseract OCR [6], the situation is improving. The Tesseract OCR, under intensive development at Google, outperforms other Open Source OCR engines in many cases. We hope the performances of these Open Source OCR engines will be improved further.

Two of the WeOCR servers are equipped with some private programs to improve the performance or to enhance the functions. Any developers can provide experimental services in this way without having their program codes open to the public.

We would like to ask more developers and researchers to deploy OCR servers to demonstrate or to show off their state of the art engines, and also to provide free services for constructing a ubiquitous OCR network to make OCR technologies popular among users worldwide.

3 A Survey on WeOCR Usage

3.1 Access Statistics

We are interested in how frequently the WeOCR servers are used. Note that every image submission page (portal site) has a warning message: “Do not send any confidential documents.” Some detailed information about the privacy is also provided.

A survey was conducted from Nov. 2005 to Sep. 2007. Figure 3 shows the monthly access counts on

the two WeOCR servers, appsv.ocrgrid and ocr1.sc, in our laboratory. The numbers of requests to six engines are added together. The line “access” represents the numbers of requests, while another one “image_ok” shows the numbers of successful image uploading. The failures are mainly due to file format mismatch and interrupts of data transmissions.

The number of accesses has been increasing gradually since the first server was put into service in 2005, despite the privacy concerns. We have about 3,000 accesses per month today. This suggests that there are a lot of applications of OCR in which privacy does not matter. The most popular engine is Ocrad, and its usage accounts for about 61% of the total number of requests within the survey period.

Unfortunately, many of the recognition results are full of garbage and are not so satisfactory. It seems many people are using the servers on a trial basis so far. Some people seem to be repeatedly using the servers after they have found some nice conditions for better recognition results. We probably need to add better OCR engines with sophisticated preprocessing, layout analysis, and postprocessing as soon as possible.

3.2 Analysis of Processed Documents

In order to analyze the applications of OCR, we have examined the images sent to the server appsv.ocrgrid, which is hosting four OCR engines. The survey was conducted from July 1 to July 31, 2007. We have manually classified 294 images by visual inspection. At least 13 images appeared to be popular test images and synthetic test data. We cannot show the actual images in this paper because of the privacy protection.

Figure 4 depicts the image types. 42.9% of the images are taken from the computer screens. This high rate is mainly because the users grabbed the

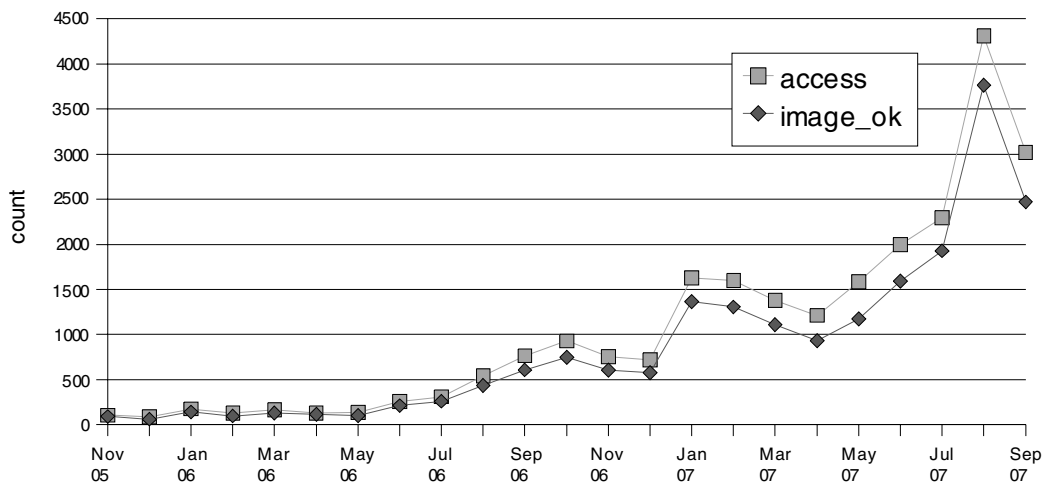


Figure 3: Monthly process counts. (appsv + ocr1)

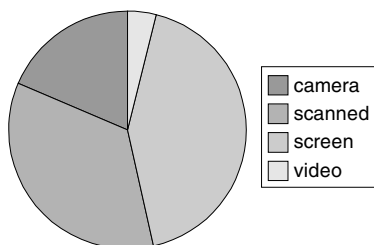


Figure 4: Types of submitted images.

Table 2: Document types of submitted images.

type	%
CAPTCHA	21.6
book	19.1
table/form	13.7
signboard	11.9
text only	10.4
scene	4.7
application screen	4.3
video (caption)	4.0
menu	2.9
comic	2.2
handout	2.2
fax invoice	1.8
journal/newspaper	1.4
others	5.8

on-screen images just for trial purposes. However, some images suggest that the users actually wanted to have some image text recognized. Some applications do not allow users to pick up text data by a pointer device.

The images taken by digital cameras accounts for 18.7%. This percentage is much higher than we expected. Most of the images are scene images. Some book images and license plate images are included. Digital camera is obviously becoming a handy document input device.



Figure 5: An example of CAPTCHA image.

Table 2 shows the breakdown of the document types. Note that the ranks are not so reliable because the sample set is quite small.

One of the popular data types is the CAPTCHA [7]. CAPTCHA is widely used today for protecting websites from robot attacks. We have seen a lot of CAPTCHA or CAPTCHA-like images (Fig.5). The users' intentions are not clear. Some people may be challenging the CAPTCHA system or the OCR engine. Some others might be just examining some text images with complex backgrounds. CAPTCHA image would not be an important target of OCR since it is designed *not to be recognized* by any OCR.

Table/form processing and signboard recognition seem to be strongly desired. The menu images were taken by a camera. Automatic recognition and translation of food menus must be very useful for travelers.

Some other interesting objects are comics, invoices, product packages, receipts, gene codes, and personal memos.

A more detailed survey using a much larger set of samples is interesting and may be included in our future work.

3.3 Applications of the WeOCR

The WeOCR platform has been used in some different ways.

Majority logic has been known to be useful for combining multiple classifiers and improving the accuracy of character recognition [8]. We have

also studied the accuracy improvement based on the majority logic using the WeOCR (OCRGrid) platform in [5]. The private OCR engines were deployed on a local WeOCR platform in our laboratory.

People in the “Seeing with Sound – The vOICe –” project have made an e-mail interface for mobile camera phones². The server receives an image from the mobile phone and send the recognized text data back to the user. An OCR engine on the WeOCR platform is used as a back-end OCR server of The vOICe server. The system was originally developed to help visually-disabled people to read text. Any OCR developers can help those people through the WeOCR platform.

One of our students has developed a virtual OCR server that can automatically select the most appropriate OCR engine from the WeOCR platform. The virtual server sends the image to multiple OCR engines concurrently, evaluates the linguistic likelihood of each recognition result, and presents the best data to the user.

We have been working on an autonomous robot system with text detection and character recognition capabilities [2]. The use of the WeOCR platform is planned.

4 Conclusions

We have introduced the WeOCR system, which was designed to realize a distributed, cooperative, synergetic OCR environment. We believe the system would be useful not only for end users worldwide but also for researchers and developers working on OCR systems.

According to the survey conducted for almost two years, the number of accesses to the WeOCR servers has been increasing gradually. The statistic suggests that many users are willing to use online OCR despite the privacy concerns. There is an urgent need to provide the users with better OCR engines.

People are actually trying to submit a wide variety of documents to the OCR servers. Although most of the document types had been predicted by researchers, we can still find some interesting objects. We may be able to obtain some useful hints for future design of OCR systems by studying the OCR usage further.

5 Acknowledgements

A part of this work was supported by the Grants-in-Aid for Scientific Research, Exploratory Research

²Seeing with Sound – The vOICe – Project : <http://www.seeingwithsound.com/>

No.18650036 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

- [1] M. Koga, R. Mine, T. Kameyama, T. Takahashi, M. Yamazaki, and T. Yamaguchi, “Camera-based Kanji OCR for Mobile-phones: Practical Issues,” in *Proceedings of 8th International Conference on Document Analysis and Recognition (ICDAR 2005)*, 2005, pp. 635–639.
- [2] M. Tanaka and H. Goto, “Autonomous Text Capturing Robot Using Improved DCT Feature and Text Tracking,” in *Proceedings of 9th International Conference on Document Analysis and Recognition (ICDAR2007) Volume II*, 2007, pp. 1178–1182.
- [3] S. Lucas and C. González, “Web-Based Deployment of Text Locating Algorithms,” in *Proceedings of First International Workshop on Camera-Based Document Analysis and Recognition*, 2005, pp. 101–107.
- [4] H. Goto and S. Kaneko, “Synergetic OCR : A Framework for Network-oriented Cooperative OCR Systems,” in *Proceedings Image and Vision Computing New Zealand 2004 (IVCNZ 2004)*, Akaroa, New Zealand, 2004, pp. 35–40.
- [5] H. Goto, “OCRGrid : A Platform for Distributed and Cooperative OCR Systems,” in *Proceedings of 18th International Conference on Pattern Recognition (ICPR2006)*, 2006, pp. 982–985.
- [6] R. Smith, “An overview of the Tesseract OCR Engine,” in *Proceedings of 9th International Conference on Document Analysis and Recognition (ICDAR2007) Volume II*, 2007, pp. 629–633.
- [7] L. von Ahn, M. Blum, N. Hopper, and J. Langford, “CAPTCHA: Using Hard AI Problems for Security,” in *Advances in Cryptology, Eurocrypt*, 2003, pp. 294–311.
- [8] H. Miyao, Y. Nakano, A. Tani, H. Tabaru, and T. Hananoi, “Printed Japanese Character Recognition Using Multiple Commercial OCRs,” *Journal of Advanced Computational Intelligence*, vol. 8, pp. 200–207, 2004.