Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

A sustainable development OCR system in CADAL application*

HUANG Chen (黄 晨)[1], ZHAO Ji-hai (赵继海)[1], HU Xiao (胡 晓)[2]

([1]*Zhejiang University Libraries, Zhejiang University, Hangzhou 310027, China*)

([2]*Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, IL 61801, USA*)

E-mail: chuang@lib.zju.edu.cn; jhzhao@lib.zju.edu.cn; xiaohu@uiuc.edu

Received Aug. 5, 2005; revision accepted Sept. 10, 2005

Abstract: This paper briefly introduces the main ideas of a sustainable development OCR system based on open architecture techniques and then describes the construction of an optical character recognition (OCR) center built on computer clusters, for the purpose of dynamically improving the recognition precision of the digitized texts of a million volumes of books produced by the China-US Million Books Digital Library (CADAL) Project. The practice of this center will provide helpful reference for other digital library projects.

Key words: Sustainable Development, Digital Library, optical character recognition (OCR), China-US Million Books Digital Library (CADAL)
doi:10.1631/jzus.2005.A1312 **Document code:** A **CLC number:** TP391

INTRODUCTION

China-US Million Books Digital Library Project is a research and development project proposed by Chinese and American scientists, aiming at creating a universally free access digital library containing over one million scanned books, using optical character recognition (OCR) whenever possible to support full text searching (http://www.cadal.cn). It is one of the key projects of the Ministry of Education of China for the Tenth Five Year Plan, and called China-America Digital Academic Library (CADAL) Project in the nation (DRC, 2004). This project is based on an open framework, and the amount of its resource will reach 50~100 Terabytes.

There are various kinds of resources included in the CADAL project, such as thesis and dissertations, block-printed editions, ancient books and other valuable traditional cultural resources. The following major difficulties exist in efficiently using these resources:

(1) Digitizing the materials with OCR is one of the key issues of the project. However, there is no available OCR software that can recognize and digitize all kinds of source materials for the project.

(2) An OCR software requires a large amount of computing capacity. In traditional methods, software applications are installed on specific platforms and manual operation is adopted. It results in not only low efficiency of software and hardware resources but also complexity in management.

(3) Considering potential further development, the scanning resolution is set as 600 dpi. However, currently common OCR cores are all developed based on B/W (black & white) images with 300 dpi. Therefore, how to use the redundant information contained in 600 dpi images to improve the precision of OCR is also a focus of the system (Chen *et al*., 2004).

(4) Since the precision of OCR software cannot reach 100%, manual collation is a necessary procedure. A method is therefore needed to provide a uniform operation procedure support and management mechanism. At the same time, we need to pay attention to the inheritability of manual work investment.

* Project supported by China-US Million Books Digital Library Project

SUSTAINABLE DEVELOPMENT PERSPECTIVE ON THE OCR SOLUTION

Currently, a common method of digitization is the mixture of OCR and manual collation (Kim *et al*., 2004). However, for a project of over one million books in multi-languages and multi-types, there are no referable examples (Shaw, 2000). We believe it is uneconomical to simply repeat the same manual work. The main idea of this paper is to create a sustainable development dynamic OCR system which will, with the development of new technologies, continuously improve text recognition precision for the whole million digital books.

Sustainable development was firstly proposed by the World Commission on Environment and Development in a report titled "Our Common Future" (Bruntland, 1987). It has obtained common understanding in the international society. The concept was defined as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs". Here we borrow it to illuminate the openness and inheritability of our system.

Openness means not to stick to one OCR core, but to flexibly take advantage of all OCR cores, adopting different cores according to different types of books. Inheritability denotes assuring system reuse: with breakthroughs in new technologies, the system can continually improve OCR precision for the whole million digital books. By repeating the process, the system can achieve high quality text corpus of over one million books with very little manual intervention.

Open system and embedded core

Sustainable development requires a system to be continually improvable, to have open interface specifications, to be stable when running software applications, and to be adaptable. Specifically, the requirements are as follows:

(1) A continually improvable OCR framework, which can inherit obtained achievement.

(2) Open system interfaces, which make it possible to embed OCR products on the market and to further improve text quality when OCR techniques progress in future.

(3) No more manual intervention than normal operators.

(4) System improvement in the aspects of both OCR and text. Improvement in OCR is the special OCR processing of 600 dpi images; improvement in text covers a broad range, such as the fine categorization of content and print quality, and the cooperation and association among them.

To sum up, how to design interface specifications is the key to realizing such an open system. We have designed a system with precise structure and flexible interfaces by using the ideas of pattern-based development and model-driven framework in software engineering (Brunelli and Writer, 2004; Sparks, 2005), adopting the design idea of virtual machines, and imitating the concept of "microkernel" in operation system (Evi and Yang, 1999). The microkernel approach consists in defining a core system with only basic scheduling functions. Other instantiation services, such as segmentation, text analysis, OCR, etc., are implemented as "peripheries" and "plug-ins". Such approach reduces the association among system components, and thus enhances the normalitiveness and openness of the system.

According to this idea, we can embed an OCR core in the application level, access OCR result data through standard communication protocols and transfer the data to post-processing modules. In this way, we can farther most assure the openness of the system.

Inheritability of manual intervention

One characteristics of OCR is that it relies highly on manual intervention during its early stages. With proper intervention, recognition precision can be greatly increased. To be brief, on the base of automatic layout analysis, properly involving manual confirmation, which distinguishes pure text content from hybrid content of table, equation, chart and text, will make the subsequent OCR more satisfactory (Bu *et al*., 2004). On the other hand, for resources of specific types, symbols and areas that will probably cause recognition error can be found in some OCR results. Manual intervention for this can also effectively assure output quality.

As a part of the system design, adding automatic recording and learning processes of manual intervention is the key to assuring sustainable development. The activities of manual intervention will be fully recorded and preserved. When the system adopts new OCR cores, or adjusts new error control procedures to

do circular recognition, the memory of layout analysis obtained by self-learning will be effectively activated, and thus a new round of OCR can output high precision result without further manual intervention.

SYSTEM MODEL CONSTRUCTION

The designed system consists of three independent software modules: data preprocessing, optimized recognition and manual intervention, which respectively correspond to three physical components: data center, data optimization center and manual intervention center. The architecture is shown in Fig.1.

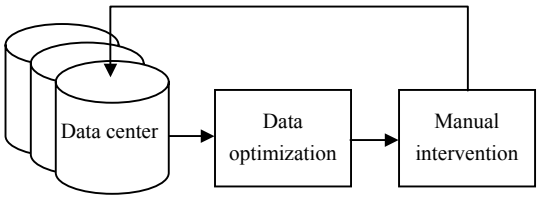

Fig.1 Operation architecture

(1) Data center is composed of high-powered cluster servers and a storage area network (SAN);

(2) Data optimization center uses high-powered cluster servers running all analysis and recognition software;

(3) Manual intervention center is accessed by special VPN methods so that commercial partners can process the data at other places.

According to the architecture described above, data optimization center is the gist of assembling and running software modules. The optimization process has the following steps: data analysis and classification, page segmentation, data OCR, result inspection and manual intervention. These steps form a system in a flow driven mode. Every module is relatively independent, and can be replaced or updated by independent development or by adopting mature commercial software. Consequently, this architecture will assure the openness of the system in the best manner.

Intelligent inspection module

In OCR practices, some steps are difficult for automatic processing, but can remarkably benefit from manual intervention. Layout analysis in Hybrid content pages and font recognition are such examples (Chen and Ding, 2004). Unlike traditional OCR systems, our architecture has an intelligent inspection module which will prompt a need for manual intervention when it detects continuous text with high error rates after analyzing originally obtained data and identifying the original resolution, as shown in Fig.2.

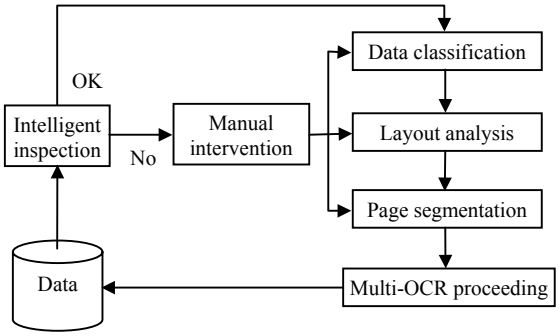

Fig.2 Optimization process flow chat

In a large-scale production environment, every page of a document is marked as an independent job. The system will conduct the OCR flow continuously and send pages with possible errors to the task pool of the manual intervention center where these pages will wait for manual intervention and then be sent to subsequent processing.

In fact, the intelligent inspection module runs through the whole optimization process. It participates in evaluations in the steps of page classification, layout analysis, segmentations, etc. The main process is described below:

(1) The quality of original texts is automatically measured by using text collation software. If the quality is good, it directly goes to the last step—automatic text analysis and correction. Otherwise, it goes to one of the two branches in Step (2) according to the precision of the original texts;

(2) If the precision of original texts is within a certain range, it will go to the segmentation step. If the precision is bad, manual intervention will be used to confirm the material categories before it goes to the segmentation step;

(3) Segmentation and segmentation integration by voting are conducted. If the segmentation effect is good, the materials will be recognized by different software corresponding to the material types, and the recognition results will be voted on, integrated and

presented. If the effect of segmentation integration is not good, manual layout analysis will be used before segmentation is done again;

(4) Recognition results are analyzed. If the result quality is good, text analysis and automatic correction will be done. Otherwise, the layout will be automatically analyzed and its effect is then measured. In the case of good effect, the layout will go to the segmentation recognition step again. If the effect is not good, the layout will go to the manual layout analysis step where the layout recognition results are manually corrected before it goes to the segmentation recognition step again;

(5) Results of automatic text analysis and correction will be directly written into databases. So far the optimization process has been accomplished.

The intelligent inspection module is involved throughout the process described above and directly affects the fashion in which the data is processed. Therefore, Fig.2 is just a sketchy demonstration which cannot fully show this module's roles in the optimization process but helps explain its functions and ways to participate in the process.

Manual intervention record module

The limited manual intervention can greatly improve at low cost the general quality of OCR. However, in order to ensure the effective operation of the open system, we must add a mechanism making the results of manual intervention continually reusable and thus maximizing its benefit. This mechanism is implemented in an intervention record module, as shown in Fig.3.

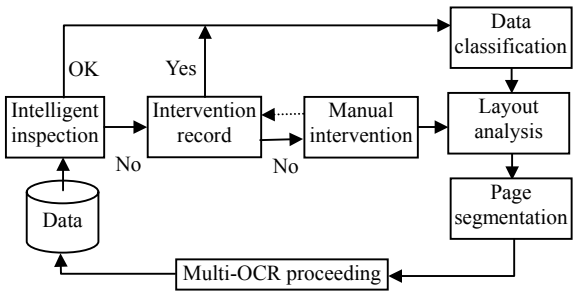

Fig.3 Successive model

The intervention records are prerequisite for assuring the system's inheritability. As described above, the openness of the system makes all functions in the process replaceable. When better applications of these function modules are available, the system automatically runs the process shown in Fig.2. During the processing, the manual intervention original results being effectively utilized is a key factor in promoting recognition quality. The functions of inheritable modules are bi-directional: on one hand, all manual interventions are automatically preserved in the intervention records; on the other hand, when the expected results output from the intelligent inspection module are not good, optimization process can be conducted using previous parameters retrieved from the intervention records. In this manner, all manual intervention will have to be conducted only once, and then to be reused within the system. This assures sustainable development and sustainable use of the system as a whole.

DISCUSSION

Currently, the CADAL administrator center has finished formal discussion on the design scheme, and a test on a small pipeline system, based on which a large-scale system will be constructed.

To a computing project with such a large amount of data, which operation framework should be adopted is a legitimate problem. Traditional OCR systems are all single computer ones. Even in professional data processing companies, a single computer running a whole OCR proceeding from layout analysis and page segmentation to character recognition, with many human operators is widely adopted, while collaborative systems are rarely used. Following this mode would hurt the system's efficiency. For example, layout analysis and segmentation occupy little computing time, but the system has to wait for the OCR process in order to continue the subsequent steps. For this reason, we separate each step from others in the OCR process shown in Fig.2, and form a combination of multiple PC groups. Fig.4 demonstrates the flow-based PC group.

In this model, computer deployment will be adjusted according to practical situations. For example, when processing complicated pages, page segmentation will need more computation, and thus we need to add more computers to do segmentation. This dynamic adjustment mechanism is implemented by a

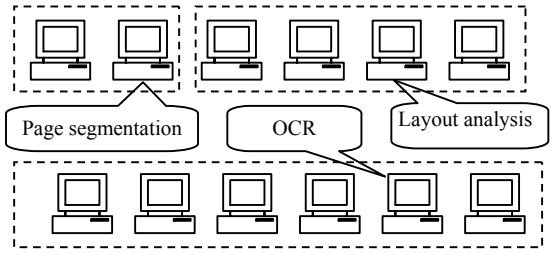

Fig.4 Flow-based PC group

module called resource scheduler. Based on the actual operating status of the system, the resource scheduler uses preprogrammed procedure to schedule the computing resources of the PC group. It also dynamically adjusts and assigns resources according to node status, and thus uses the PC group in a highly efficient manner (Fig.5).

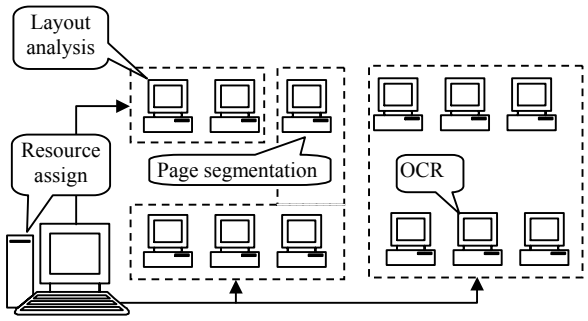

Fig.5 Dynamic adjustment model

The openness of the system requires the integration of existing OCR cores. There are two ways for such integration: one is to purchase commercial OCR SDK and to do second development; the other is to directly use existing commercial OCR software. For the above design scheme, the SDK second development is better for ensuring system efficiency. However, the SDKs of mature OCR software producers, like FineReader of the United States and Tsinghua Wintone of China, are very expensive and have limit on the user's licenses. So, it is hard to be applied to a completely open system. As a result, we turn to the second approach and need to consider how to effectively integrate the results of commercial OCR software. We have succeeded in some OCR software, and obtained achievements in phrases of the project. We can also input intermediate results of the optimization process so as to get final recognition data. However, there have been no proper technical solution models to parse all OCR software. Therefore, the openness of this design scheme is still a focus of researchers' attention.

CONCLUSION

The OCR project in CADAL is totally different from traditional OCR applications. It emphasizes continuously improving the text precision of a very large collection of data. We proposed effective solutions in the aspects of system structure, implementation steps, and data reuse. Based on the accomplishment of the full-text data required by the CADAL project, we hope to continually improve the recognition precision of electronic books through the independent operation of the system, and to implement a dynamic digital library model that will eventually increase the legibility and availability of books. We also expect the practice of this design scheme will provide helpful reference for other digital library projects. The proposed system can also be used to serve other projects, realizing the goal of sustainable development.

ACKNOWLEDGMENT

The authors would like to thank their colleagues, Zhen Yan, Chen Haiyin, He Liang, and Jin Genda of the Repository and Service Team in CADAL for their enormous contributions to the reported work. Many thanks also go to Dr. Zhang Yuzhi from the Computer Department of Chinese Academy of Sciences and Leon Zhou from IBM Grid Computing, ISG, for their advice and design for the system. Thanks also go to Yang Yu, Fan Xiangdong from the Datum Data Company for their efforts related to testing and implementing the OCR proceeding.

References

Brunelli, M., Writer, N., 2004. The Holy Grail of Model-driven Development. http://searchwebservices.techtarget.com/qna/0,289202,sid26_gci999474,00.html.

Bruntland, G.(Ed.), 1987. Our Common Future: The World Commission on Environment and Development. Oxford University Press, Oxford.

Bu, F.Y., Liu, C.S., Ding, X.Q., 2004. Distinguish tables from

graphics in layout analysis. *Computer Engineering and Application*, **12**:83-87.

Chen, L., Ding, X.Q., 2004. Font recognition of single Chinese character based on wavelet feature. *Acta Electronica Sinica*, **32**(2):177-180.

Chen, Y., Sun, Y.F., Zhang, Y.Z., 2004. A study on segmentation method for gray document image. *Journal of Chinese Information Processing*, **18**(4):44-49.

DCR (Development and Reform Committee), 2004. The Approval for Report on Results of Feasibility Study on Construction Project of the Chinese Academy Digital Library & Information System (CADLIS)'s Tenth Five-Year Plan Authorized by Development and Reform Committee, China, No. 2004-1649 (in Chinese).

Evi, N., Yang, J.Z.H., 1999. UNIX System Administration Handbook. Tsinghua University Press, Beijing.

Kim, M.S., Ryu, S., Cho, K,T., Rhee, T.H., Choi, H.I., Kim, J.H., 2004. Recognition-based Digitalization of Korean Historical Archives. Asia Information Retrieval Symposium AIRS 2004. Revised Selected Papers (*Lecture Notes in Computer Science*, **3411**:281-288).

Shaw, E.J., 2000. Building a digital library: a technology manager's point of view. *The Journal of Academic Librarianship*, **26**(6):394-398.

Sparks, G., 2005. MDA Overview. Sparx Systems. http://www.sparxsystems.com/bin/MDA%20Tool.pdf.