第五屆文字識別與文件分析研討會
Jan. 1996.

ROSE：a Rapid OCR Software Development Environment

I.S.Shyu(徐英士), *Y.N.Lien(連耀南), M.Y.Chen(陳謀琰), W.W.Lin(林文雯),
L.T.Tu(屠榮梃), R.Y.Tsay(蔡榮儀), M.C.Liang(梁明智), Y.S.Huang(黃雅軒)
Computer & Communication Research Laboratories
Industrial Technology Research Institute
W300/CCL, 4F #11, Park Ave II, Science Based
Industrial Park, Hsinchu, Taiwan R.O.C.
Tel: 886-35-781390 Ext. 306
Fax: 886-35-781398
E-mail:shyu@OCR2.ccl.itri.org.tw

Abstract

ROSE (Rapid OCR Software Development Environment) is a software development environment consisting of a library of basic reusable OCR software modules in C programming language、a training database、a testing database、and a parameterized Shell level command suite for various design tasks. It is intended to facilitate rapid OCR software construction process in meeting the rising demand of shorter OCR development cycle under limited resource constraints. Given an OCR product specification, software developers need only to determine a set of parameters such as which characters set in which category to be used、which feature extraction method to be used、and which matching algorithm to be used. They can then quickly construct a feature database by merely executing a Shell script with determined parameters, and test the algorithm they designed at Shell command level. This fast prototyping process facilitates OCR designers to design the best performance OCR empirically. Finally, an efficient OCR software with product strength can be constructed based on the parameters chosen in the design phase by "plug-in" the software modules in the C level library.

1. Introduction

Character Recognition has been studied for almost twenty years in Taiwan and is still a challenging problem today. All three different input types: printed、on-line、and handwritten, and various languages including Chinese、alphanumerals、and Japanese were tried to be recognized by many research laboratories and universities. Some native products for recognizing printed and on-line Chinese and alphanumeric characters have appeared in the markets for many years.

The state of arts in designing and developing an OCR system is summarized as follows: [1]

(1) There are two schools of recognition philosophy: statistical and is structural. They both have advantages and weakness: the former performs better in absorbing noises and deformation but worse in discriminating similar character categories. Some researchers are working on combining both statistical and structural approach in serial or parallel manner to get optimal results.

(2) Recognition process consists of two phases: training phase and recognition phase. In training phase, representing sample data are collected and used to analyze the average properties of the target characters. Discriminating imaginary features of each character are extracted as the referenced features of that character. In recognition phase, imaginary features are extracted from the images of those unknown characters and are compared with the reference features of sample characters to find the best match.

(3) Recognition process consists of two major subprocesses: feature extraction and classification. A general document OCR system may also include image processing、layout understanding、segmentation、recognition、and post processing subprocesses.

(4) OCR is an empirical science. There is no theoretical、mathematical or universal solution to recognize different types and kinds of characters. The researchers are always trying to find new features、new combination of features、new matching techniques、new character image processing methods、new

*Yao-Nan Lien is now an Associated Professor of Department of Computer Science, National Chengchi University.

distance functions、new clustering algorithms and many other new ways to improve recognition rates.

(5) The importance of reliability are highlighted in recent years when applying OCR technology to mission critical tasks because of the following two reasons:

a. great liability in legally or monetarily related applications.

b. manual error correcting will eat out the cost saved by the automatic recognition. Lately A "multi-expert" concept is popular in OCR field to challenge this tough issue.

(6) The specification given to an OCR design usually consists of the character set、recognition speed、recognition rate and reliability. These objectives usually conflicts to each other so that the challenge of design is usually not on the high recognition rate itself, but on the balance of these conflicting objectives.

Because of reasons mentioned above, a time-consuming iterative design process is required to find an "optimal design" for a given specification. On the other hand, because of the advance of computer and communication technologies, the pace toward a global information society is greatly accelerated recently by the explosion of Internet related applications such as Gopher and WWW. One of the greatest challenge to achieve that goal is to digitalized the huge amount of information accumulated in the human history. Therefore, there is a great need to accelerated the process of OCR system design in facing the foreseeable demands [2]. Actually, we have been seeing examples where the slowness of OCR system design hurt the customer's expectations [3].

ROSE is designed to meet that expectation. It consists of a library of basic reusable OCR software modules in C programming language [4]、a training database、a testing database、and a parameterized Shell level command suite for various design tasks. It is intended to facilitate rapid OCR software construction process in meeting the rising demand of shorter OCR design and development cycle under limited resource constraints. Given an OCR product specification, software developers need only to determine a set of parameters such as which characters set in which category to be used、which feature extraction method to be used、and which matching algorithm to be used. They can then quickly construct a feature database by merely executing a Shell script with determined parameters, and test the algorithm they designed at Shell command level. This fast prototyping process facilitates OCR designers to design the best performance OCR empirically. Finally, an efficient OCR software with product strength can be constructed based on the parameters chosen in the design phase by "plug-in" the software modules in the C level library.

To be more specificially, ROSE is constructed on a UNIX operating system thanking to the open characteristic of UNIX. It has the following features:

(1) Major software components such as software architecture、training sample database、and configuration management are organized according to their semantic structure. This semantic structure plays an important role to maintain a structural consistency among all software modules with minimum compromise.

(2) Large amount of different real-case document and character images for training and testing are collected.

(3) A universal naming convention is created to enforce the naming of all software objects including files、variables、samples、etc. This naming convention is designed to be consistent with the semantic structure.

(4) All software modules are managed by a thorough configuration management compliant with the ISO-9000 standard. To simplified the versioning, each development version is controlled by a set of environmental variables. By setting proper environmental variables, a software developer can extract proper version of software modules from the universal software repository.

(5) It has a set of Shell commands available for fast prototyping. A designer can quickly construct an OCR system by writing a Shell script with proper parameter specified. A prototyping design cycle is greatly shorten from weeks or months into days.

(6) It has a set of C-level reusable library containing all major software components. Efficient OCR software can be quickly constructed from this library.

In section 2 we will describe the Configuration Structure in more detail. In section 3 the standard of ROSE file names and formats are proposed. In section 4 the Shell commands are introduced. In section 5 we introduce the construction of a four-level OCR

library. At last, in section 6 we make some concluding remarks.

2. Semantic Structure

The ROSE semantic structure consists of four parts: (1)databases (2)OCR core(3)Applications and (4)User Interface.

2.1 Databases

All databases used for training and testing an OCR system are stored in directory DB. Directory DB stores the following three subdirectories: Document 、 Character 、 and ChinesePhrase. The former two databases are both decomposed into Image and Description parts. Document images come from four sources: CEDAR 、 CENPARMI 、 IPTP and CCL. The first database consists of handwritten alphanumeric characters and words extracted from dead envelopes collected by the United States Post Office. The CCL database consists of images scanned from magazines 、 books 、 newspapers 、 tax forms 、 fax forms 、 checks 、 envelopes 、 etc.. Some of the documents are printed materials, while some of them consist of handwritten as well as mixed types of characters. Most of these documents were collected from our business clients. (Refer to Fig. 1)

To get the design parameter and evaluate the performance when developing and testing a document analysis algorithm, a "correct answer" concerning the layout 、 line numbers 、 field numbers 、 block numbers 、 character numbers and character categories 、 etc. of each document should be given before the algorithm is developed. Such "correct answers" are stored in subdirectories Description.

As to character images, it consists of subdirectories Printed and Handwriting. The former consists of ChineseComplex 、 ChineseSimplex 、 Alphanumeric 、 Japanese and Korea characters. The latter consists of on-line and optically scanned handwritten ChineseComplex 、 alphanumeric 、 and Japanese characters. Extracted features 、 clustering templates and matching templates are stored in subdirectories Feature.

Subdirectory Chinese Phrase stores general post process database and post address phrase database.

2.2 OCR core

Directory OCRcore consists of subdirectories of LayoutUnderstand 、 Segmentation 、 Recognition 、 Postprocessing and Utility. Subdirectory LayoutUnderstand stores algorithms developed for understanding the layout of envelopes 、 forms 、 and printed documents. All the algorithms developed for finding the minimum bonding boxes of each character in this document will be located at subdirectory Segmentation. Subdirectory Recognition consists of subdirectories storing utilities and functions made for character image processing 、 feature extraction 、 feature matching and code transformation. The integrated recognition engine will be stored in subjectory RcgFrame for application needs as well as for easily making a multi-expert classifiers. Subjectory MultiExpert stores the developed algorithms which combine various recognition engines and give high accuracy recognition results. (Refer to Fig. 2)

Subdirectory PostProcess stores algorithms which use phrase knowledge to increase character recognition rate. These knowledge consist of Chinese phrase as well as check sum or check digits in daily-used alphanumeric string. (i.e. citizen identification numbers)

At last subdirectory Utility stores utility functions considering transporting the OCR algorithms to different platforms for various users' needs. Right now only UNIX 、 Win32 and Win16 are targeted.

2.3 Applications

Subdirectory Applications consists of various products that integrate software components in the subdirectories under OCRcore, including PCCR (Printed Chinese Character Recognition) 、 HCCR (Handwritten Chinese Character Recognition) 、 FormReader 、 etc..

2.4 User Interface

Subdirectory UI consists of the following three subdirectories: Win16 、 Win32 、 and Motif. Each subdirectory stores displaying 、 editing 、 and other operating utilities implemented by C 、 visual C 、 or visual basic. They are used when executing the OCR system under different operating systems.

3. Standard ROSE file names and formats

Six files are typically used for designing an OCR recognition engine:

(1)reference character set
(2)character image
(3)character feature
(4)matching template
(5)clustering template
(6)candidate and matching distance

Designers use reference character set (1) to identify the character categories belonging to an application character set. He then extract features from these character images (2) and generate reference feature files (3). Two reference template: matching template (4) and clustering template (5) are made from the features extracted from training samples, one for recognition and the other for classification. By classifying features we get a candidate and matching distance file (6) of each character category. The information contained in this file can be used to improve the recognition system.

Since there are many design automation mechanisms built into ROSE, all data objects mentioned must be properly named according to the naming convention in order to make those design automation work properly.

3.1 Naming Convention

3.1.1 Character Image

In order to make file names valid in both UNIX and DOS file system, each file name consists of two parts: primary name and extended name. The primary name must have two characters and five numerical digits. The first character represents the character type. The second digit represents its language type. The third to the seventh digits represent the serial number of Big-5 code for Chinese characters or JIS code for Japanese, and ASCII codes for alphanumeric characters. An extended name must have 3 numerical digits, which represents the sample set number which the character category belonging to.

For example, pc00154.001 contains the first sample set of a printed Chinese character image whose big-5 serial number is 154(thus obviously belonging to daily used character set). As to another example, hj00234.002 contains the second sample set of a handwritten Japanese character image whose JIS serial number is 234.

3.1.2 Feature

The primary file name consists of the name of the feature and the primary file name of the character image. The extended file name are FTR in abbreviation of "Feature". For example, PBApc00154.001.FTR contains feature values extracted by Peripheral Background Area (PBA) feature from the first sample set of a printed Chinese character whose big-5 code serial number is 154.

3.1.3 Reference Character Set 、 Clustering Template 、 Matching Template 、 candidate and matching distance

These files are correlated because the latter three files contain information extracted from the character set registered in the first file. Their relationships are registered in the file header of each file. The extended file names for these files are IDX 、 CLS 、 and TPL in abbreviation of "Index" 、 "Cluster" and "Template". The extended file name of Candidate and distance file is DST in abbreviation of "Distance".

3.2 File Formats

Each ROSE file is composed of three parts 、 i.e. Common header 、 File header 、 and Data. Information about magic number 、 version number 、 data type and file identification string are registered in Common header. An example of the Common header is shown in Table 1.

Byte 0~3	Byte 4	Byte 5	Byte 6~11
magic	version	data_type	file_id
char	char	char	char
CLW3	0~9,A~F	1 ASCII 3 UNIX_BIN	CCxxx? ECxxx? DCxxx? SCxxx? JCxxx? KCxxx?

Table 1 : Common header of a valid ROSE file

4. Command Language for Parameterized OCR Design

ROSE provides a set of parameterized Shell commands to facilitate rapid OCR engine prototyping. Considering in the stage of Feature-Extraction, the experienced designers may change image segmentation method and cut numbers. He may extract a few set of features in different dimensions and combine them by different normalization and weighting functions. As to the classification stage he may combine different feature sets as the clustering features. He may also combine different feature sets as matching features when making template matching during different matching level. Even the matching distance functions could be changed during different matching level. Since so many design parameters could be infinitely changed in the OCR design process, a well-documented 、 ease for use 、 flexible and efficient tools should be constructed to decrease the frequency of modifying and compiling 、 linking the modified programs which usually leads to a very long design cycle.

To greatly shorten the design cycle, ROSE provides a set of Shell commands each performing a primitive OCR function. After determine the design parameters, a designer can quickly construct an OCR engine by combining these Shell commands into a simple Shell script. For example, the following "exf" command performs a feature extraction with following parameters:

```
exf  -F SDF -x 8 -y 8 -I App1.IDX
     -D $Home/DB/path/Image
```

By this command, designers will easily understand that it's SDF (Stroke Density Function) feature to be extracted ; the image segmentation will be 8 by 8 ; the character set will be App1.IDX ; and the character images can be found at $Home/DB/path/Image.

Right now a few commands are implemented, including:

- exf -- for feature extraction
- template -- for making matching template
- kmean -- for making clustering template
- distance -- for measuring the distance between samples and reference template
- psort -- partial sorting for quickly find the candidates
- imgsort -- sort the character sample images by quality
- imgshow -- to show the bit image of characters on the screen

5. OCR Library

ROSE offers a C-level library which contains many reusable C functions which can be combined to construct an efficient OCR engine with industrial strength. It is divided into four abstraction levels as shown in Fig. 3.

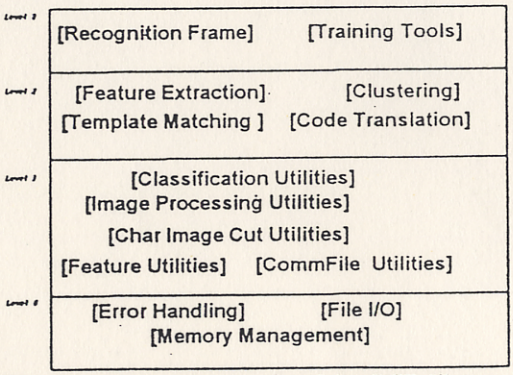

Fig. 3 : Abstraction levels of the OCR library

where each level plays a different role :
(1) Level 0 consists of device level common utilities. These utilities are physical system dependent which must be rewritten when porting an OCR system to different platforms. Since all system dependent functions are grouped together in this level, it provides physical system transparency to all other levels.
(2) level 1 consists of recognition utilities which are used in reading/writing valid ROSE files 、 transforming image formats 、 segmenting character images 、 extracting basic

features、and sorting matching results. It is a tool box which contains optional tools to help building an OCR system.

(3) level 2 consists of feature extraction functions、template matching functions、clustering functions、and code translation functions. The interface of each function is independent of any recognition target. Each function contains both data and program, where the data are grouped into the same data structure and can be processed only when the caller use ROSE read/write utilities to process the data.

(4) level 3 consists of the integrated recognition engines and tools which help to design the OCR system. The recognition engine could be directly used by other system integrators.

To sum up, this four-level architecture library:

- is easy to port to other platforms for its separation of physical level (level 0) and logical level utilities;
- is easy to reuse for its rich set of common recognition and feature extraction utilities;
- is easy to integrate into an OCR engine for the availability of highly integrated template frames.

Furthermore, this library is organized into ROSE configuration management system such that it is very easy to manage.

6. Conclusions

This paper proposes a software development environment to facilitate rapid OCR system design and development. After analyzing the design processes of various statistic OCR systems, we create a semantic structure to enforce the structural consistency among all software modules with minimum compromise. All software modules are organized and managed according to ISO-9000 standard. It offers a Shell level command suite to facilitate parameterized OCR engine prototyping. A typical design cycle can be shorten from weeks or months to days by writing a simple Shell script with proper parameter settings. It also offers a C-level library which contains many reusable C functions which can be combined to construct an efficient OCR engine with industrial strength.

By using ROSE we have constructed two recognition engines. Character Image databases used for these two experiments are CCL Handwritten Chinese Character Image database and CCL Handwritten Digit Character Image database. The recognition rate is 90% for HCCR (Handwritten Chinese Character Recognition) and 99% recognition rate for HDCR (Handwritten Digit Character Recognition).

7. Acknowledgments

The work reported in this paper is supported by the Ministry of Economics, R.O.C. The project code is 35N7100.

8. References

1. Lo-Ting Tu, "Introduction of OCR Research and System Design," CCL Technical Journal, Nov., 1992.
2. A. Amiri, A.C.Downton, S.J.Hanlon, C.G.Leedham,S.M.Lucas and D.Monger, "OSCAR:a visual programming toolkit for offline handwritten forms recognition," IWFHR-IV,pp.441~448,1995.
3. Yao-Nan Lien, Ing-Shyh Shyu, Lo-Ting Tu, Mou-Yen Chen, and Win-Win Lin, "Design of FR1000 Form Reading System," Proceedings of the First Workshop on Real-time and Media Systems, July 1995, pp. 293-302
4. Mou-Yen Chen, Win-Win Lin et al., "A Chinese Character Recognition Library for CCL Form Reader System," CCL technical report, Feb. 1994.

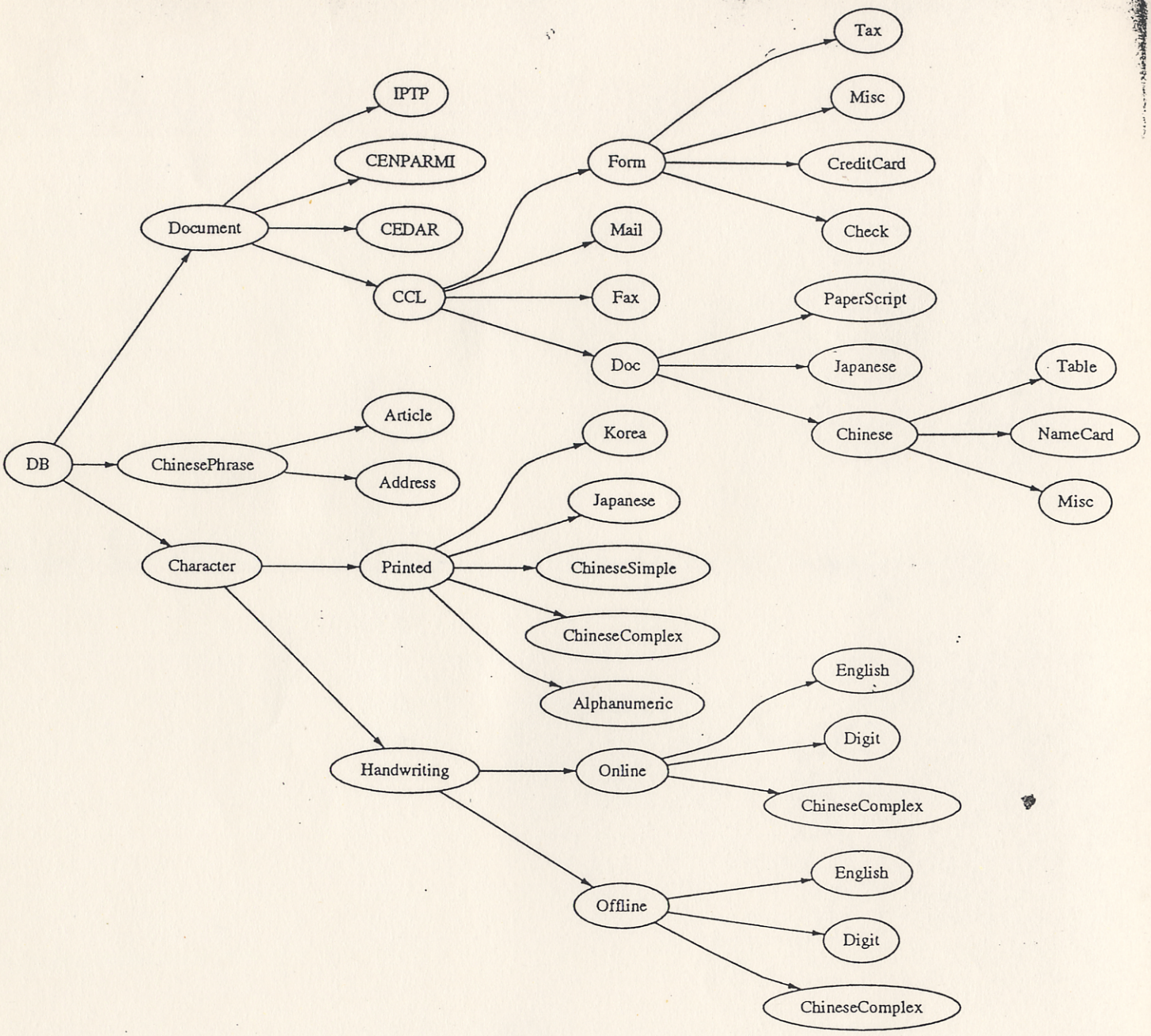

Fig.1 : ROSE Configuration structure of Databases

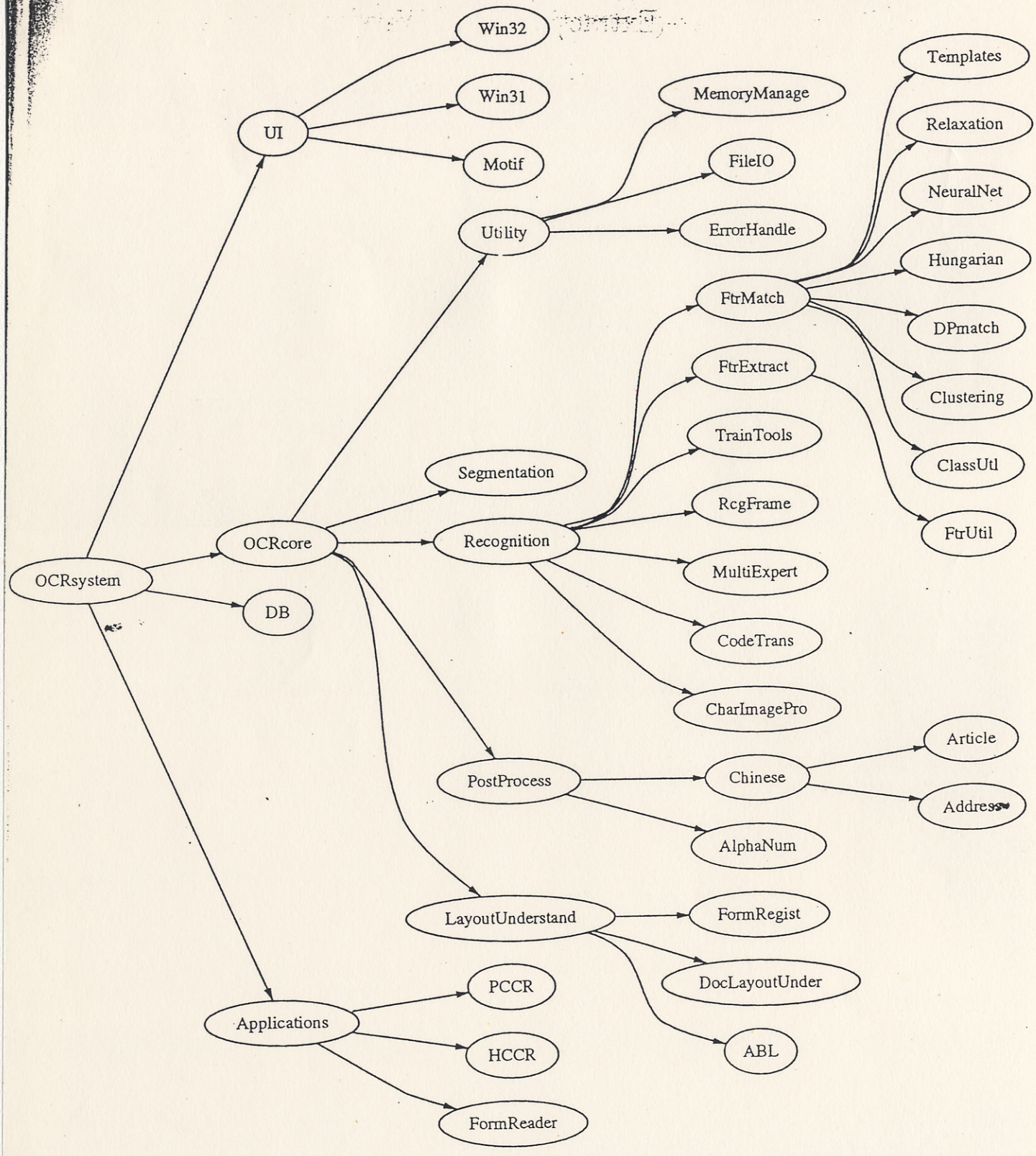

Fig.2 : ROSE Configuration structure of OCR system