



Deliverable Report

Ref.: DR_FRESH_WP2.5

Vers.: DR_FRESH_1.0

Date : 14/02/07

Page : 1 /5

Client : European Commission

Project : FRESH

Project N°: FP6-516059

Project Number:	FP6-516059
Document number:	DR_FRESH_WP2.3.3_fd18
Document Title:	Final delivery
Document status:	Final
Date:	14/02/07
Availability:	Confidential
Authors:	AlgoTech Informatique

Abstract	This report summarizes the work done on DICTIONARY (Structure, semantic, I/O processing)
----------	---

Keyword List	WP2, dictionary, text recognition.
--------------	------------------------------------



Deliverable Report

Ref.: DR_FRESH_WP2.5

Vers.: DR_FRESH_1.0

Date : 14/02/07

Page : 2 /5

Client : European Commission

Project : FRESH

Project N°: FP6-516059

DICTIONARY (Structure, semantic, I/O processing)

1. INTRODUCTION	3
2. STRUCTURE	3
2.1. ADDING NEW VALUES	3
2.2. GLUED CHARACTERS	3
2.3. RECOGNITION RATE.....	3
3. CORRECTION PROCESS	4
3.1. CONFUSED LETTERS AND NUMBERS / UPPER CASE AND LOWER CASE	4
3.2. WORDS.....	4
3.3. STRUCTURES.....	4
4. INTERFACE	4



DICTIONARY (Structure, semantic, I/O processing)

1. INTRODUCTION

After the text recognition, where each character is recognized without taking into account its position in a word, some incorrect results can occur because of two main reasons.

- The bad quality of the source document induced by storage or duplication conditions.
- The presence of glued characters which are considered as single characters.

In these cases, using a dictionary can improve the global recognition rate of a document.

2. STRUCTURE

In the technical documents, the text is mainly extracted from symbols parameters. The spelling rules are not matching with literary style.

Many symbol names are very long, and are terminated by numbers, or make reference to the name of the manufactured product.

Many parameter values are represented with physical units and are terminated by a name which can be stored in a list.

2.1. ADDING NEW VALUES

Because of the text specificity previously defined, it is necessary to be able to complete the contents of the dictionary and then adapt it to every new particular case.

2.2. GLUED CHARACTERS

The calculation of the number of characters, evaluated from the bitmap representation of the word to recognize, can give a hint on the presence of glued characters.

This calculation can use the height of each isolated block into the word image.

After evaluation of the mean number of characters, the dictionary can be filtered according to a range of sizes of words.

A confidence rate is then calculated for every dictionary entries and the best result is proposed to the user.

2.3. RECOGNITION RATE

The recognition process evaluates several words at a time. The term “word” refers to a whole block identified during segmentation. For each word, the recognition system classifies it into several probable



Deliverable Report

Ref.: DR_FRESH_WP2.5

Vers.: DR_FRESH_1.0

Date : 14/02/07

Page : 4 /5

Client : European Commission

Project : FRESH

Project N°: FP6-516059

character sequences, and gives each a confidence rate. This “combination” is a method inherent to the classification technique.

Each image corresponds to a segmentation block, which is then classified to a word. The probable words are then used to match against dictionary entries. High probability words can be discarded if no suitable dictionary entry is found.

3. CORRECTION PROCESS

3.1. CONFUSED LETTERS AND NUMBERS / UPPER CASE AND LOWER CASE

A first correction can be done for each alphabetic character which has a representation similar to that of a numerical character, or which has a representation similar to that of capital letter and small letter.

The indication of the relative size of each character in the word (number of pixels in the source picture) and of the presence or not of characters among which the distinction between the capital letter and the small letter is sure, can allow a first correction.

3.2. WORDS

The text layer of an electrical drawing does not only consist of symbol parameters, although this constitutes the major part of the text to be recognized.

There is also some free text, which corresponds to a technical vocabulary concerning the functions represented or the names of manufactured products.

3.3. STRUCTURES

Widely represented in an electrical drawing, they are generally relative to the symbol parameters. They are often characterized by a set of alphabetical and numerical characters.

These terms cannot be listed in an exhaustive way. It is therefore necessary to associate them to a generic coding.

4. INTERFACE

Before the text recognition, the operator can choose the level of use of the dictionary.

- Correction of capital letters, small letters and confusions between alphabetical and numerical characters.
- Correction with the technical vocabulary registered in the dictionary.
- Correction with the generic structures corresponding to the syntax of the symbols parameters.
- Correction with typical sentences.

The introduction of a new sentence in the dictionary leads the addition of the words and the structures which it contains. So the maintenance is easy and fast.

