



## Deliverable Report

Ref.: DR\_FRESH\_WP2.3.2  
Vers.: DR\_FRESH\_1.0  
Date : 13/10/06  
Page : 1 /6

Client : European Commission

Project : FRESH

Project N°: FP6-516059

Project Number:	FP6-516059
Document number:	DL_FRESH_WP2_2.3.2_Final
Document Title:	<b>Processing tool and method for human intervention</b>
Date:	3rd July, 2006
Abstract	This document summarizes the recommendations for user interaction in the recognition process.
Keyword List	WP2, Identification and Recognition of Textual Parts



## Deliverable Report

Ref.: DR\_FRESH\_WP2.3.2  
Vers.: DR\_FRESH\_1.0  
Date : 13/10/06  
Page : 2 /6

Client : European Commission

Project : FRESH

Project N°: FP6-516059

# Contents

1 Introduction .....	3
2 Recommendations on user interaction for error correction at character level .....	4
2.1 OCR error correction recommendations .....	4
2.2 Dictionary .....	6
3 Recommendations on user interaction for selection of re-training data.....	6



## Deliverable Report

Client : European Commission

Project : FRESH

Project N°: FP6-516059

# 1 Introduction

Task 2.3 of workpackage 2 deals with the identification and recognition of textual parts. It describes the proposed method for text recognition and the reader is referred to this deliverable for all details on the recognition method, as it is the main document for this task.

In the FRESH project proposal, it was written that the whole tool will be interfaced with an interactive validation mode, to give the user the possibility to correct wrongly recognized text.

Deliverable D.2.3.1 deals with:

User feedback is usually given under the form of a choice of a correct classification of a character/word. In fact, this is basically one sample (character → classification) that can be added to the training set. The system may deal with this additional sample in several ways:

- *Re-training*: Re-do all the training according to the new sample.
- *Inner correction*: This is a process that searches for the source of the mistake and tries to eliminate it.
- *Direct application*: Usually the user will be asked for feedback when there was a doubt in the classification. The system can directly apply the user choice to other situations where the same doubts have appeared.

When incorporating user feedback, some problems may occur:

- *Over-training the algorithm*: This is the problem of obtaining an algorithm that is too much trained to work perfectly on a particular set of characters, that it will loose accuracy on other sets.
- *User mistakes*: Users also make mistakes, which may add noise to the training.

These problems are not specific of user feedback, and may also occur in any training set used. However, they can be very well analysed “in laboratory”, and corrective measures may be taken in this case. But when the system is “on production” we must be very careful because we loose control on the training samples that the user feedback is adding in real-time.



## 2 Recommendations on user interaction for error correction at character level

In classical OCR, a page of text is usually scanned and submitted to the OCR module; after that post-processing is performed through dictionary lookup and spell-checking. After that, non-recognized words are presented to the user as they are in a usual spell-checker and the user can interact to correct recognition. There is usually no re-training in commercial OCR packages.

Such an approach is difficult to extend to the problem which deal with in the Fresh project. The data are scarce compared to the extensive training which can be performed for mainstream OCR. But the classes of characters to be recognized are also few, mainly digits and capitalized letters.

Therefore, we recommend that the user interaction for error correction at character level be performed in the following way:

1. All subimages of characters recognized as belonging to the same class are presented to the user in a single line of table for quick inspection.

The user can then quickly point out those characters wrongly recognized and immediately, through a drop-down menu, select the correct recognition.

This presentation will be made on a separate screen, without referring back to the original image.

The corrections made by the user will be memorized throughout the session and it will be possible to save them for another session.

2. After spell checking using the specific “grammar” or set of regular expressions defining the valid “words” to be found in a wiring diagram, the system will highlight in a specific color, this time directly in the original image, the “words” which are not recognized as being valid according to the specification.

The user will thus be allowed to directly type the correct word, or to validate the recognized word if it was just a limitation of the grammar or specification.

### 2.1 OCR error correction recommendations

In front of each alphabetic character, all the subimages of therecognized characters are listed. The current version of the Algo'tech software doesn't yet use the recognition score to order the results, but after having selected an image, all the similar images in the line or in the whole

grid can be automatically selected for comparison. In this way, the same correction can be quickly and efficiently applied to several characters (Fig. 1).

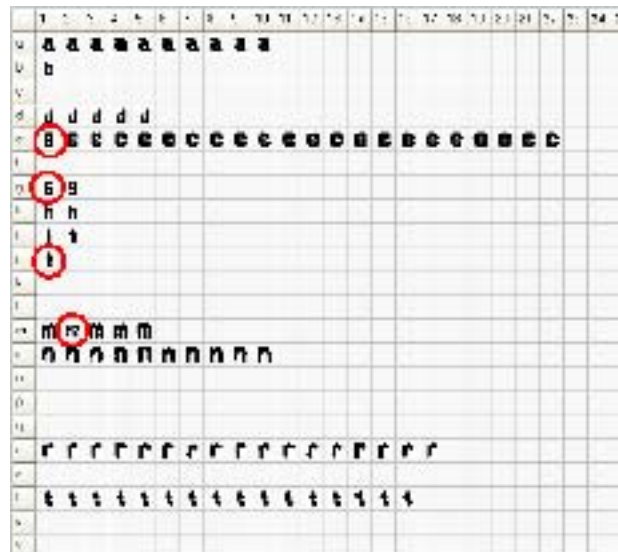


Fig. 1. OCR correction interface.

When some images are selected in the grid, the words in which there are included in the original image are presented to the user. Some characters are difficult to differentiate, such as for example '8' and 'B' (Fig 2).

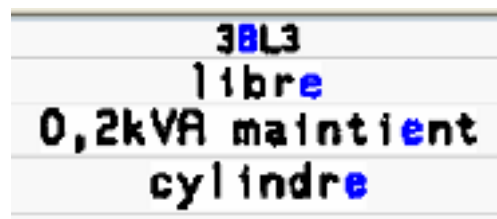
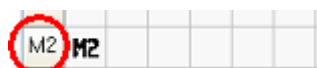


Fig. 2. Characters difficult to differentiate.

Moreover, when an image represents connected characters, a new line is added to the grid where all the same configurations will be affected. Indeed, when a special configuration of connected characters occurs in the original image, it is usual to find the same configuration several times (Fig. 3).





## Deliverable Report

Ref.: DR\_FRESH\_WP2.3.2  
Vers.: DR\_FRESH\_1.0  
Date : 13/10/06  
Page : 6 /6

Client : European Commission

Project : FRESH

Project N°: FP6-516059

Fig. 3. Same configuration detected.

### 2.2 Dictionary

The specific grammar used into electrical drawings is encoded in the dictionary with generic characters such as `\*`, `!`, `?`.

### 3 Recommendations on user interaction for selection of re-training data

Among the mis-recognized characters in the first step, the user may want to select some for re-training of the recognizer.

Therefore, we recommend that at the end of a session with the recognizer, the user be presented with all mis-recognized and corrected characters and asked to select those he wants to add to training data for the recognizer.

These data are then added to the training data and re-training is run, after warning the user about the danger of degrading the classifier if selecting bad training data.

The user will always have the choice to backtrack to the last instances of the classifier, or to get back to the initial settings.

An initial set is to be automatically provided to user so that he can choose this “automatic” additional training set if he doesn’t have the availability to carefully elaborate this set. This automatic set shall be built starting from all of the corrected characters, and discarding those that produce the most impact on the training set balance.

The user will have the ability to save the re-trained versions of the classifier in different files or directories. When working on a particular page, he/she will be able to choose the version of the classifier to use. It shall be possible also to re-train already re-trained versions of the algorithm.