



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 1 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

Project Number: FP6-516059
Document number: FRESH-D.2.3.1-Text-Recognition-V3.0-FINAL.doc
Document Title: Final Report on Text Recognition
Document status: Draft
Date: 30/06/06
Availability:
Authors: Pedro VIEIRA (Tekever)

Abstract **This document presents the final report on Task 2.3 – Identification and Recognition of Textual Parts.**

Keyword List Optical Character Recognition



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 2 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

Table of contents

1. INTRODUCTION	3
1.1. PURPOSE.....	3
2. BACKGROUND	4
2.1. OCR PROCESS STEPS.....	4
2.2. OCR MAIN ISSUES	5
2.2.1. SEGMENTATION	5
2.2.2. FEATURE EXTRACTION	5
2.2.3. RECOGNITION.....	7
2.2.4. COMBINATION OF CLASSIFIERS	7
2.2.5. HANDWRITING OCR	9
2.2.6. APPLYING SPECIFIC DOMAIN KNOWLEDGE	9
2.2.7. PERFORMANCE MEASURES.....	12
2.2.8. APPLICATIONS.....	12
2.3. OCR TOOLS	12
3. APPLICABILITY IN FRESH	15
3.1. FRESH SPECIFIC PROBLEMS	15
3.2. RESEARCH FOCUS.....	16
3.2.1. ALGOTECH OCR ALGORITHM EVALUATION.....	16
3.2.2. USING FEATURES AS INPUTS TO NEURAL NETWORKS.....	17
3.2.3. ALTERNATIVE CLASSIFICATION METHODOLOGIES.....	18
3.2.4. USING USER FEEDBACK IN THE LEARNING PROCESS	18
3.3. COMBINATION OF COMMERCIAL OCR PACKAGES	19
3.4. ADDITIONAL NECESSARY RESEARCH	20
4. RESEARCH PERFORMED	21
4.1. APPROACH	21
4.2. DATASETS	21
4.3. SINGLE CHARACTER RECOGNITION	21
4.3.1. BATCH TEST METHODOLOGY	21
4.3.2. FEATURES.....	22
4.3.3. CLASSIFICATION ALGORITHMS.....	22
4.4. CHARACTER SEGMENTATION.....	24
4.5. ADDITIONAL INFO	27
4.6. CONCLUDING REMARKS	27



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 3 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

1. INTRODUCTION

1.1. PURPOSE

This document presents the results of the FRESH Task 2.3 – Texts recognition.

A study of state-of-the-art OCR (Optical Character Recognition) methods which are candidates to be used within the FRESH project, was performed. This was based on a review of several up-to-date papers and paper abstracts which were published in several conferences, journals and books related with OCR. Their references are presented below.

Some of the techniques were implemented and tested, using typical datasets of the domain (electrical diagrams provided by EADS-Sogerma). Further research was performed under this task, in order to adapt the techniques to the FRESH specific problem, and to further improve the results. Summaries of results are presented when applicable.

The remainder of the document is structured the following way: Section 2 presents a background on OCR, including a basic structure for the OCR process, the latest developments in OCR main issues and some OCR commercial. A study of the applicability of the techniques to FRESH is presented in section 3 (including an evaluation of the AlgoTech OCR algorithms). Finally, a description of the research performed up to the moment is presented in section 4.

2. BACKGROUND

2.1. OCR PROCESS STEPS

OCR is a typical Pattern Recognition problem. Specifically, it can be sequentially classified as Pattern Recognition → Image Pattern recognition → Document Recognition → OCR. From there, one can divide the problem further into Character Recognition or Word Recognition.

[Koerich00] proposes the OCR process architecture in Figure 1:

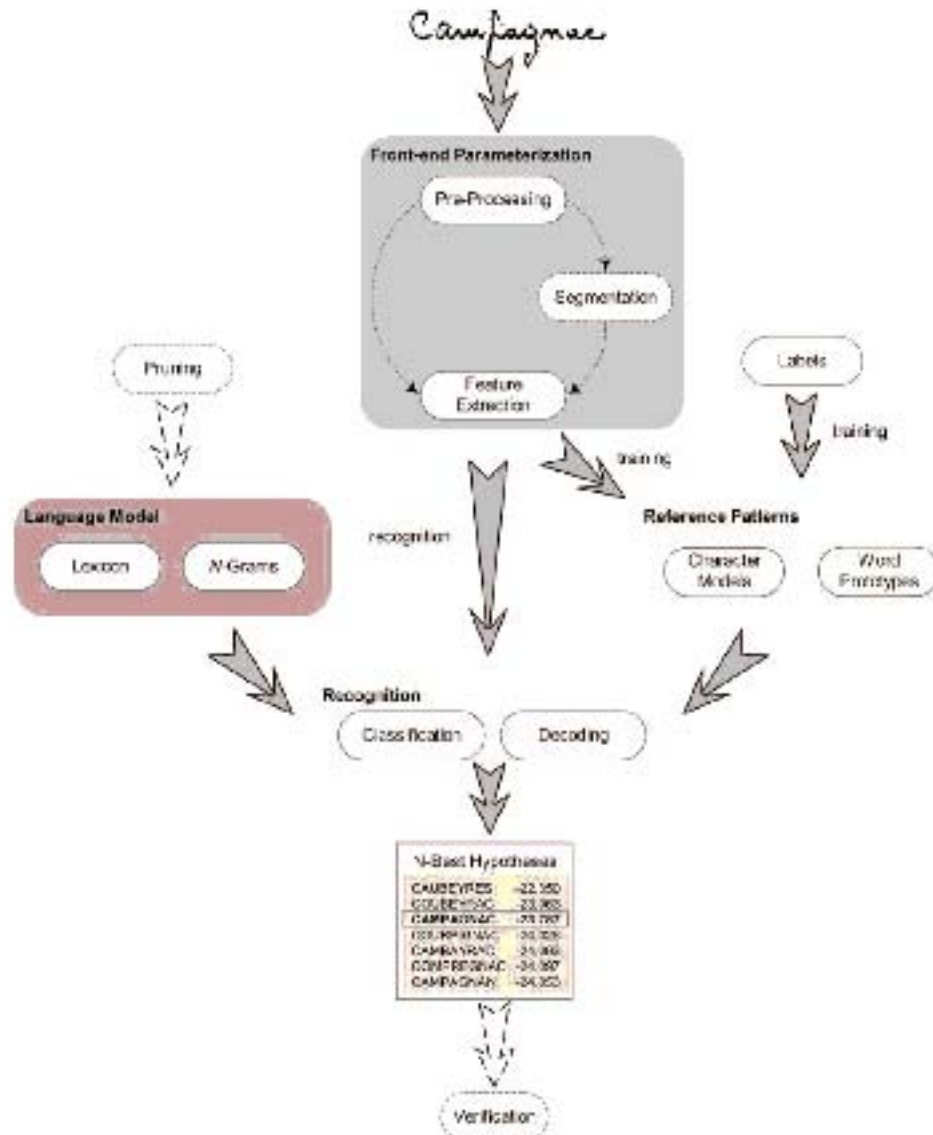


Figure 1 – OCR generic process [Koerich03]



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 5 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

Note that in the above process it is not considered the preceding step of text-region extraction: this would be the pre-step where images corresponding to text lines would be extracted from a document image, and isolated into separate system inputs.

Although the process consists of apparently independent modules, the fact is that the applied methods are highly dependent on each other. Next, we describe the most crucial steps and the implications of their results in the whole process:

- **Pre-processing** – The process from which the original scanned image is transformed into a binary image.
- **Segmentation** – The process from which a text region is separated in text lines, words, sub-words and characters.
- **Normalisation** – A set of operations performed at image level in order to reduce the variability of patterns.
- **Feature extraction** – the process of converting an input image to a set of descriptive values with some meaning (features).
- **Recognition** – a methodology both to store information on the specific problem (through training) and applying this information to transform the image of text into a character string (classification).

2.2. OCR MAIN ISSUES

A multitude of methods can be applied in each step of the OCR process. The possibilities are vast, and almost all known techniques of Pattern Recognition have been applied to a given OCR problem. This section introduces the main issues in OCR research and how they are dealt with in literature.

2.2.1. SEGMENTATION

Regarding segmentation, two types of OCR systems exist:

- Segmentation-based
- Segmentation-free

Segmentation-based systems are systems that depend on the segmentation process to separate the words into characters, and then recognise these characters separately.

Segmentation-free systems are systems that intend to recognise words, instead of characters, without including a word-to-character segmentation process.

Segmentation-free OCR is preferred when the amount of words is small and/or the segmentation process is difficult.

2.2.2. FEATURE EXTRACTION

Feature extraction is the process of converting an input image to a set of descriptive values with some meaning. The biggest problem is the choice of descriptive-enough features.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 6 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

The total set of features to choose is enormous, and can be virtually any algorithm that transforms an image to some descriptive model of it, be it a vector of numeric values vectors, or some higher-level description.

There is no specific set of features which have proved to be the best for OCR problems. However, the following types of features have received more attention in literature [Koerich03]:

- Low-level features:
 - Smoothed traces of the character contour
 - Stroke direction distributions
 - Pieces of strokes between anchor points
 - local shape templates



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 7 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

- Medium-level features (that aggregate low-level features to serve as primitives):
 - edges
 - endpoints
 - concavities
 - diagonal and horizontal strokes
- High-level features
 - Ascenders
 - Descenders
 - loops
 - dots
 - holes
 - t-bars

Research has also been made in Feature Selection methodologies. This is the approach in which the feature set is automatically selected by the machine-based algorithm. The idea is to provide an initial big set of features, from where the algorithm identifies the most descriptive ones. An example of such algorithm is the Floating Forward Feature Selection algorithm (see Appendix B).

2.2.3. RECOGNITION

Next, we describe the most adequate techniques that OCR research has pointed out, regarding recognition.

Neural Networks: Neural Networks are definitely the preferred approach for recognisers, in cases of small variability of patterns.

Word-based recognisers: For more complicated problems, word-based recognisers are used. These imply the existence of a lexicon (dictionary) of possible words, besides the simple list of characters.

Also, **rejection mechanisms** may be applied to discard uncertain or unlikely recognition results (usually delegating the decision to the user).

2.2.4. COMBINATION OF CLASSIFIERS

Apart from pure research on recognition techniques, most non-trivial applications of OCR are based on the combination of several classifiers. The basic idea is simple: the most voted result, or the one that achieves the highest confidence value, regarding the weighted outputs of the several classifiers, is chosen.

Ensemble Methods is a technique based on the definition of several recognition techniques that are applied to a given problem, both during training and classification.

Boosting techniques are automatic methods to discover the best way to combine a set of classifiers, mostly regarding the discovering of weights, through the analysis of classifiers performances.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 8 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

These techniques also apply to the combination of complete OCR commercial packages, in a similar manner.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 9 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

2.2.5. HANDWRITING OCR

Two main additional challenges are introduced in Handwritten OCR:

- The variability of patterns is tremendously augmented, both in shape and size, (decision regions are hard to define in an optimal way).
- Characters get connected with each other in many different manners, making character segmentation much more difficult.
-

Handwritten character recognition requires:

- The adoption of a word-based recognition methodology
- A more complete lexicon
- Normalisation before feature extraction
- As much domain-specific information as possible (such as context information).

2.2.6. APPLYING SPECIFIC DOMAIN KNOWLEDGE

Domain knowledge appears mostly in the following formats:

- Lexicons
- Context
- Structural pattern recognition methods
- Hierarchical pattern classification

2.2.6.1. Lexicons

For handwritten character/word recognition and, in general, to cope with noise in the input text images, a **lexicon** (dictionary) of possible words (or character combinations - Ex: digit-letter-digit-digit-{N|P}...) is often used to improve the recogniser accuracy. This is actually the only possible improving solution most of the times, becoming indispensable in the worst cases.

There are two ways to include a lexicon in a recognition process:

- Matching the recognition results with the dictionary words (or grammar rules), and obtaining the most similar word. This is the simplest solution, which is used mainly when the classifier performance is rather good, becoming the lexicon a mere tool for validation and ambiguity resolution.
- Involving the lexicon in the recognition process, using dynamic programming methodologies or Hidden Markov Models. In this case, word-based recognition is applied, and the lexicon becomes a structural description of the word patterns. This is adequate to handwritten OCR, or when character segmentation is a difficult process.

When the lexicon to be used is composed of a set of words or rules that may change, dependant on context, we are in the presence of a **dynamic lexicon**. Dynamic lexicons are also characterised by the fact that not all the words that compose the lexicon are present during training. This means that segmentation-free techniques are not feasible in these cases.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 10 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

2.2.6.2. Context

The idea behind context information use is to have some sort of description of the problem that is being treated, in order to tailor the recognition algorithms to specific issues. Limiting/choosing the lexicon to apply is also a result from context dynamism application.

Several types of context dynamism may be applicable:

- Document context: The type of document/page that is being scanned implies a specific OCR strategy.
- Image context: location in the image/page may provide information on expectable problems to solve.
- Structural context: the relation of a given text region to other objects in the image may give an indication of the type of problem to be attacked.

2.2.6.3. Structural Pattern Recognition

Structural Pattern Recognition is a paradigm by which the patterns are described by a hierarchical structure, and divided into components.

Word-based recognition is a form of structural pattern recognition, in the sense that a word is described as a sequence of characters, either statically or according to rules of character composition.

Structural pattern recognition is a form of implicating some domain knowledge in the process, since the hierarchical structure of patterns is specific to the domain. To deal with this model, several methods exist.

2.2.6.4. Hierarchical pattern classification

A hierarchical classification of patterns is also a form of applying domain knowledge. We can inform the OCR system that we expect characters "B" and "8" to be similar, therefore belonging to the same higher-level group. The global idea is to perform top-down discrimination on a tree-like hierarchy of character groups. On each node of the tree, a different classifier is applied, distinguishing only between a small set of patterns.

This method has the following advantages:

- Simplification of the recognisers between groups in the hierarchy
- Simple application of different recognition strategies
- Allows the recognition of highly variable patterns (through a step-by-step part-elimination strategy).

...and disadvantages:

- Applicable only to well-segmented patterns.
- Involves user-intensive training and problem solving
- Not robust (new notations and new specific problems always require further analysis).



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 11 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

2.2.6.5. On-line Adaptation

All recognition algorithms need some sort of learning mechanism to train the parameters for a given application. The idea of on-line adaptation is to continue the process of learning while using the system on-the-job. For instance, make use of the user responses in doubtful situations to further tune the recogniser parameters. This is a form of on-line real-time application of domain knowledge.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 12 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

2.2.7. PERFORMANCE MEASURES

The following performance measures are used in general OCR studies:

- Classification accuracy
- Sensitivity to sample size (how does system accuracy changes with the size of the training set);
- Ambiguity rejection efficiency (how good is the system in rejecting unknown, unclear and wrong classification results);
- Outlier resistance (how robust is the system to the presence of objects that do not correspond to its intended OCR application).

2.2.8. APPLICATIONS

The following applications are the ones that have received more attention in recent OCR literature:

- Bank checks
- Postal address read
- Digital libraries (Document indexing and search)
- Handwritten numerals (generically) → Many applications to digit recognition (smaller pattern set).
- Mathematical expressions recognition
- Car license plate recognition

2.3. OCR TOOLS

Some COTS OCR tools were identified. Some belong to the group of the most used and efficient tools in the market. Others were identified as interesting concepts worth evaluation. However, no evaluation of these tools was possible yet. Some of the providers of these tools request a pre-acquisition agreement before providing a trial version.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 13 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

FORMATION Dossier n°.		OPTS		
A01	DOSSIER EXEMPLE	NG125L		
Q2		C16		
4-6 3A		Puissance	5x(1x300) / PH	
7BA			3x(1x300) PELN	
NOV04-5			LOCAL G.E.	
AIRCRAFT WIRING MANUAL		2 GROUPES ELECTROGENES 2X1600KVA		
L50	100.1	DECLENCKEUR	R15	
R11	04-19	DEFENSELNAGE	C11	
43	KM1	LAMPES	C10	
45	04-5		C9	
46	X2	9161-CF24	R11	
R14	10	9160-CF24	R11	
PM4	RACK:	9159-CF24	R10	
PAZ		shunt alarme	9158-CF24	R10
		& a carte	102RH	R9
		DIR-C105.2	R9	

Figure 2 – Difficult texts for recognition



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 14 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

FORMATION Dossiern°: QPTS
NG125L
A01 DOSSIER EXEMPLE C16
Q2
4-6.3A Puissance
78A 5x(1x500) / PH N0104-5 3x(1x300) PEN
LOCAL G.E.
AIRCRAFT WIRING MANUAL
2 GROUPES
ELECTROGENES
2(1 600K V \bar{A}
100.1 DECLENCHEUR R15
1.50 04-19 DESENFUMAGE
CII
Rlf LAMPES cio
43 KM1 c
04-5
45 9161—cFzq 6H
46 X2 9160—CF2q
R714 10 9159—CFZq 610
9158 CFZ
RACK: RIO
PAZ IOZRH
shuntfri(eme DIRtIDS. 2
alacarto

Figure 3 – Recognition results with Microsoft Office Document Imaging, OCR module, from input on Figure 2. – Errors are marked in yellow.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 15 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

3. APPLICABILITY IN FRESH

3.1. FRESH SPECIFIC PROBLEMS

The main problems to be solved in FRESH are described next. These are the ones we believe can be solved neither with common techniques nor by using COTS OCR. The problems are:

- **Connected characters:** characters frequently appear connected with each other and with other objects in the image.
- **Variable patterns:** there are fonts with different sizes, symbol shapes, and aspects. There are also handwritten characters to be recognised.
- **Wavy text lines** which often appear.
- **Bold characters**, implying: connected characters, noise, and modifications in character morphology.
- **Thin characters**, leading to missing portions of the characters.

We propose the following steps to attack the OCR problem in FRESH.

- Define a clear **structure of context** of the documents to be scanned. This shall be used not only to validate OCR results but also to choose the best applicable OCR techniques. The goal is to answer the question: "is it feasible to rely on context-oriented methods, and lexicons, in FRESH?"
- Choose an **OCR strategy** for each case:
 - Define the applicable **lexicon**, as complete as possible.
 - In contexts where small lexicons can be applied, the best approach is to use a word recognition system. Also to use word-based recognition for situations of connected text. **Dynamic programming** and **Hidden Markov Models** are candidate techniques to be researched and tested.
 - Use **segmentation-based** methods in situations where characters can be easily segmented. **Neural networks** and **Statistical approaches** are best suited to these cases.
 - For each case, study the set of most descriptive features using a battery of features and feature selection algorithms, such as **Principal Component Analysis**.
 - For visibly highly variable patterns apply techniques of **structural pattern recognition**.
- **Combine** several of the defined strategies, when the context definition is not enough to limit the type of problem to be solved.
- Consider using **commercial OCR packages** for simpler situations, or simply as starting points for the OCR process.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 16 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

3.2. RESEARCH FOCUS

3.2.1. ALGOTECH OCR ALGORITHM EVALUATION

3.2.1.1. Main characteristics

In view of the above concepts, the AlgoTech algorithms present the following main characteristics:



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 17 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

Pre-processing and segmentation

- The reports do not mention pre-processing and segmentation activities. We therefore assume that the described methods are applicable only to segmented characters.

Recognition

- The used recognition methodology was a single **neural network**. This is a technique which has shown to be robust to noise and pattern variability. Nevertheless, literature has concluded that this type of recognition scheme is directly applicable only to well segmented characters.
- A **hierarchical classification** of characters, with a two-step classification tree, to cope with pattern variability, obtaining better results.

Feature extraction

- The authors first tried to apply **directional features** (8 directions, pixel count) with poor results in what regards robustness to noise.
- The best tested feature set was a resolution transformation (in a grey-scale space) to a **reduced grid**.

Domain knowledge

- A **lexicon** of words and character sequence coherence was used, improving the results.

3.2.1.2. Reusability in FRESH

The work described on the AlgoTech reports seems to be able to be re-used for FRESH in the following ways:

- As a black-box OCR application, to be combined with other OCR packages, in a combined-classifier scheme.
- As the base for an OCR system to be applied to well separated characters, becoming one possible strategy in some contexts. Improvements might be necessary, in order to cope with more variable patterns.

3.2.2. USING FEATURES AS INPUTS TO NEURAL NETWORKS

Incorporating features in Neural Networks is not a trivial process. While the number of pixels is huge, features normally go down to less than 30. Also, 2 adjacent pixels represent spatially related information, while features are often representing more abstract and independent concepts.

To use features, instead of pixel values, a completely new neural network would have to be designed. Also, there should be less correlation between the input variables, leading to a high non-linear output function for the outputs. Adding more inner layers to the Neural Network would probably be a solution.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 18 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

3.2.3. ALTERNATIVE CLASSIFICATION METHODOLOGIES

3.2.3.1. Kohonen maps

Self-Organising Kohonen Maps are a form of unsupervised learning (information about the correct classification of samples is not given to the learning algorithm). The algorithm works into organising a set of training samples into a set of regions of similarity, according to the feature space.

3.2.3.2. Support Vector Machines

Multi-class support vector machines are applicable to classification problems where it makes sense to have classes represented by a small set of examples.

In OCR, the problem is in dealing with the variability of the shapes of the characters, given its pixels. The performance of the algorithm in FRESH is to be studied, especially in what regards the high variability of patterns and the noise coming from segmentation problems.

3.2.4. USING USER FEEDBACK IN THE LEARNING PROCESS

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. This is not problematic to recognition algorithms whose training is based on incremental processes.
- **Inner correction:** This is a process that searches for the source of the mistake and tries to eliminate it. White box classification approaches may somehow incorporate the possibility for inner correction.
- **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. This would be a specific process, independent from the main recognition algorithm.

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.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 19 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

3.3. COMBINATION OF COMMERCIAL OCR PACKAGES

COTS OCR such as the ones define in chapter 0 is to be considered an option for FRESH, due to the following reasons:

- Classification rate in printed characters is high
- Several features for manual and automatic use are present in their API's.
- Some COTS OCR packages include additional capabilities.
- Several methods exist that make use of COTS OCR in more sophisticated systems, which means that additional improvements and specificities of the FRESH project may be built on-top COTS OCR.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 20 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

The drawbacks are:

- “OCR vendors seldom publish technical details of their proprietary OCR engines”
- Post-processing issues, such as text-image registration may be insufficient, in terms of accuracy. In FRESH, this is definitely a key issue: associating the text with the corresponding image region.
- It is necessary to align text regions (knowing: which text corresponds to which, in different APIs)
- Programmatically, OCR APIs raise some problems:
 - They may fail in run-time.
 - They may include bugs, which are impossible to solve.
 - Different APIs may be difficult to integrate.

3.4. ADDITIONAL NECESSARY RESEARCH

The following additional improvements are foreseen to be necessary, either to COTS OCR packages or common techniques for OCR, mainly to improve accuracy and reliability:

- Improvements on pre-processing
- Combination of multiple classifiers
- Applying specific domain knowledge

In FRESH, 100% accuracy is necessary before storage, since an error in a digit or letter in a component name or document/region title can have tragic consequences. Therefore, further research is necessary on **methods to deal with OCR errors** in FRESH. It is expectable that:

- Some sort of human intervention will always be necessary.
- The system will support the user in the identification of the text locations that might need intervention.

Further investigation should also be made on the possibility of using the OCR results on other tasks of the segmentation/recognition process. **Interdependency** between the OCR module and other symbol recognition systems should be carefully studied.

It is foreseeable that a separate **pre-processing** step will be necessary for the text regions.

Combination of different systems should be applied in two ways:

- Different systems working together for the same place portion of text ;
- Selection of different systems for different portions of text, depending on context.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 21 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

4. RESEARCH PERFORMED

4.1. APPROACH

Given the enormous amount of different techniques to attack the problem of OCR, it was decided that the research should focus on the following aspects:

- Single character recognition → dealing with the research on **good features** for character recognition as well as a few strategies for character **classification**.
- Character segmentation → focused on algorithms for **separating connected characters**
- Word recognition → consisting on research on algorithms for **whole word recognition**, rather than 2-step segmentation/recognition.
- Inclusion of additional information → be it **user feedback**, available **dictionaries** or preliminary results on **symbol recognition**..

4.2. DATASETS

We have used several different datasets, however not very populated ones. The results obtained were mainly indicative of the power of the techniques, and do not constitute real use test cases. Additional research will have to be made, with the same techniques, but with big enough datasets.

4.3. SINGLE CHARACTER RECOGNITION

4.3.1. **BATCH TEST METHODOLOGY**

A test methodology was defined, together with a set of tools, to be able to perform huge sequences of tests with different algorithms, configurations and datasets. The set of defined tools were based on a simple command line interface, using data files on text format and a specific directory structure.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 22 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

4.3.2. FEATURES

The following features were implemented:

- FXYCenter
- FXYDeviation
- FXYprojAreaOverMean
- FXYSkewness
- FXYSymmetry
- FXYTransitions
- *Sekeleton Features*
- *Moments*

These were tested using the Bayes classifier. The Bayes classifier is one of the simplest classifiers, using probabilities of statistics performed on training datasets.

The features were calculated for all characters in the datasets. Several tests were performed, in two different perspectives:

1. Training and testing on the same dataset: This gives an idea of the discrimination power of the features in that dataset.
2. Training and testing on different datasets: This gives an idea of the robustness of features.

The main conclusion here is that the group of selected features is very different when using different datasets. Even if good results are obtained on the training sets, this reveals a lack of robustness of the chosen features when using FFFS as the feature selection method.

4.3.3. CLASSIFICATION ALGORITHMS

Although the feature selection methods were based on the classification accuracy of the Bayes algorithm, more powerful classification methods should be used to cope with the variability of patterns and lack of robustness of the selected features.

4.3.3.1. Neural Networks

Neural Networks are a common strategy for OCR, and we have started with a proven configuration for the FRESH domain, as detailed in 3.2.1- Algotech OCR algorithm evaluation. In this case, the inputs to the network are the pixels of input image, after being transformed to an n-by-n size grid.

We tried to improve results with the inclusion of features, but as explained in 3.2.2 this is not trivial. We have tested with the feature values as inputs to the neural network, with and without the grid values, but the resulting NN would have too many inputs, becoming extremely complex. With the small datasets we have used, it was just not possible to train the networks to a satisfactory result.

An alternative which improved the first results was to use not the feature values as inputs, but the accuracy of the Bayes classification for each class, in the selected feature space . This is illustrated in the following figure:

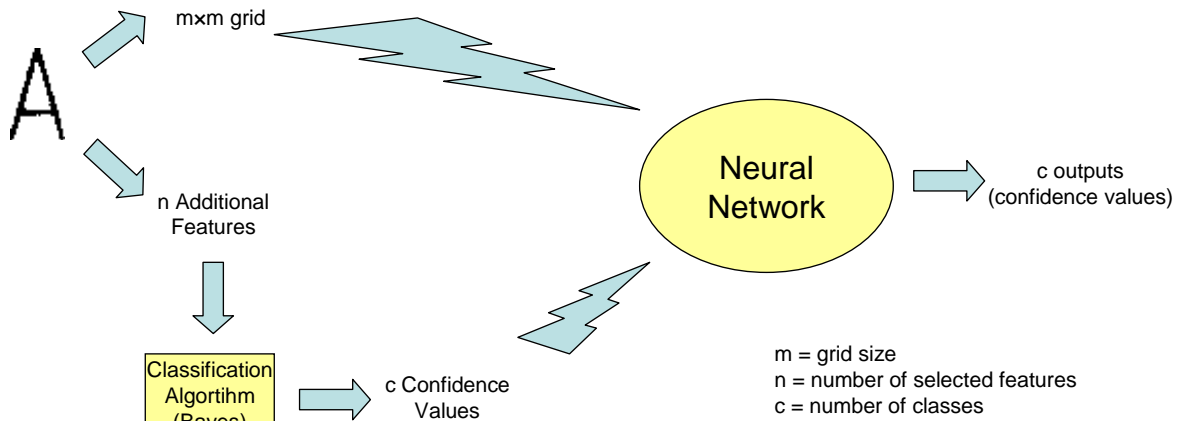


Figure 4 – Neural Network proposed scheme

The philosophy behind this configuration is the following: each pixel on the grid produces some sort of activation of a given output. Such complex relationships between input values and output values are very difficult to model with a neural network.

However, if we use a parametric classification algorithm such as Bayes, the confidence values for the outputs will have the adequate configuration.

This means that if one input slightly corrupted, the network will still be able to output a correct result, by weighing all the other inputs.

Therefore, this approach seems not only a good way to deal with the bad classification cases of the Bayes algorithm, but also a firm strategy for combining the basic neural network architecture with feature values and other classification algorithms.

We have obtained the following results:

Neural Network Configuration	Result on training dataset	Result on test dataset (see Figure 2)
<ul style="list-style-type: none"> - 64 inputs (8x8 character grid) - 62 outputs (1 per character) - 1 hidden layer w/ 64 neurons 	98%	30%
<ul style="list-style-type: none"> - 103 inputs: <ul style="list-style-type: none"> o 64 inputs (8x8 character grid) o 39 features, 1 per output class (Bayes classification) - 39 outputs - 1 hidden layer w/ 64 neurons 	100 %	46 %



Table 1 – Results on Neural Networks classification

In this case, the test dataset used was a very problematic one, as seen on Figure 2. The poor results obtained in this dataset are mainly due to the lack of heterogeneity of the training dataset used. We still need to perform more tests on bigger training datasets. However we can see that the configuration with extra inputs has performed better.

4.3.3.2. ANN (Approximate Nearest Neighbours)

ANN is a Nearest Neighbours algorithm in essence. The idea is to gain efficiency by compromising accuracy on the nearest neighbour found.

We have tested thoroughly the ANN algorithm, using Zernike Moments orders ≤ 13 as features.

When training on data similar to the test data, results can go as high as 92%. The 8% error is mostly related to the same misclassification problems, which arise from 2 main sources:

- Very similar characters types. This could be improved by using specific features for such peculiar cases.
- Invariance of the Zernike Moments to rotation and symmetry. This could be improved by adding some non-invariant features.

4.4. CHARACTER SEGMENTATION

We have identified on literature a generic process for character segmentation, or dissection, which generally follows 5 steps:

1. Filtering (on the original image)
2. Projections / Profiles
3. Filtering (on projections/profiles)
4. Peaks/valleys
5. Cut

Typically, these steps are performed only on images which were previously identified as candidates for containing more than one character.

Also, when a perfect separation of characters is not to be expected, the use of recognition results can help in the process:

- One idea is to use the above strategies to over-segment the characters and use dynamic programming for finding the best fit: segments \rightarrow character.
- Another hypothesis is to have the above strategies provide several probable segment sets, and then pick the one that yields higher recognition confidence.

We have performed some tests in the trained datasets, which are presented in the figures below. We have used:



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 25 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

1. Filtering (on the original image) – We have tried using the skeleton and contour of the image, with non-satisfactory results, so we present next results without any filters
2. Projections / Profiles – Simple vertical projection, upper profile, lower profile and the second derivative of the projection over its value.
3. Filtering (on projections/profiles) – No filters were applied at this step.
4. Peaks/valleys – We have calculated peaks and valleys on each of the projections.
5. Cut – Simple vertical cut was performed



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 26 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

	Good	Over segmentation	Bad
Projection (pixel count) - valleys (pink lines)			
Upper profile – valleys (pink lines)			
Lower profile – peaks (green lines)			
2nd Derivative – peaks (green lines)			

Table 2 – Examples of good and bad character segmentation results

The table above shows some examples of good and bad results. In each, the image of two connected characters is presented. Blue pixels override the characters, constituting the projections or profiles calculated. Pink pixels show candidate cutting lines corresponding to local valleys of the projections. Green pixels show candidate cutting lines corresponding to local peaks of the projections. In each case we are considering either only peaks or only valleys of the projection, as stated on the left.

- A “good” result occurs when there is only cutting line that correctly separates the two characters.
- An “Over segmentation” result occurs when there is more than one line segmenting the characters, with a maximum of one line cutting each character.

A bad result occurs when there is more than one line cutting inside at least one of the characters.

- The resulting method will be highly dependant on recognition results, due to the so many over-segmentation results found. Dynamic programming may be used for such.



Deliverable Report

Ref.: ID_FRESH_TEK

Vers.: version1

Status : Draft1

Date : 13/10/06

Page : 27 /27

Client : European Commission

Project : FRESH

Project N°: FP6-516059

- Combining several projection/profiles techniques may be useful, if the bad segmentation results in one technique coincide with good results in other techniques.

More research could be made on filters and cutting mechanisms. However we believe that for the given dataset, the focus should be on the aid of recognition results, with whole word recognition methods.

4.5. ADDITIONAL INFO

We divide additional info in three fields:

- Aid from user
- Lexicons/dictionaries
- Symbol recognition results

The three "sources" of information will provide the system with masks for possible words, in a given context. These masks could have the form of regular expressions or similar, limiting the range of possible characters in a given position. The output of the recognition will be separated by several rated hypotheses. The better match between the two will constitute the final result of the system.

Initial studies showed the inexistence of usable lexicons or dictionaries for the set of diagrams to be studied. This should be studied in more detail.

4.6. CONCLUDING REMARKS

We have encountered many problems on effectively training recognition algorithms, due to the small datasets we have used. However, the results obtained are already important in guiding further research with bigger datasets.

We have showed the power of some features, but need to do some more research on feature selection strategies. Here the biggest problem is finding robust-enough feature sets.

Up to the moment, we have revised a Neural Network architecture, showing some improvements on adding features to the inputs of the network. However, the best results were obtaining using the ANN classifier, becoming the preferred choice.

We have defined a method for Character segmentation (dissection), which will guide Word Recognition based strategies, with the support of Dynamic Programming. Results are promising.

We have defined the strategy for the inclusion of additional information to support the results.