A PROLONGATION-BASED APPROACH FOR RECOGNIZING CUT CHARACTERS

Antoine Luijkx, Celine Thillou and Bernard Gosselin
aluijkx@ibelgique.com
celine.thillou@tcts.fpms.ac.be
bernard.gosselin@tcts.fpms.ac.be
Faculte Polytechnique de Mons
TCTS Laboratory

Abstract

This work has to do with Optical Character Recognition in the context of degraded characters, and more particularly in the context of cut characters. It appears like a complement to a more general problem and investigates a field not often seen in literature. The problem is first divided into two sub-problems: first, modular networks, separating the detection of cut characters and their classification are proposed. Then, the classification part is completed with information coming from the prolongation of cut characters.

Keywords: Optical Character Recognition, Modular Neural Networks, Cut Characters

1. Introduction

Although the current techniques of OCR work satisfactorily, the recognition rates fall down for degraded characters. This paper presents results for the particular problem of one of common degradations: cut characters. Cut characters can be defined as incomplete characters, whose the missing part is not known. Examples of cut characters are shown in Figure 1.

2. State of the Art

Only a few papers are available on this subject. Nijhuis, in [4], suggests a technique for an automatic recognition of car license plate where, pending on the position of the camera, cut characters appear in the picture to be analyzed. Nijhuis has developed a technique based on modular networks. First, characters to be classified are sent to a neural network, previously trained on normal characters. If the outputs of this neural network are below a predefined threshold (i.e. if the confidence levels are too low), the corresponding characters

Figure 1. Example of cut characters.

are redirected to a dedicated network, built on cut characters. The results of the two networks are then compared and the final class determined. The final recognition rates obtained are very high. However, this technique only deals with characters cut on the up side, and assumes that a threshold can be defined for confidence levels.

In [2], Fukishima suggests a technique to recognize (and restore) occluded characters (character partly hidden by stains, objects, ...). The technique is based on a particular network (called neocognitron). This network contains different layers extracting particular features. One layer is dedicated to detect the occluded parts of the character and inhibits features coming from occluded zones, because they are irrelevant. The Fukishima's algorithm seems to be a viable technique, although no recognition rate is detailed in his papers, nor any discussion about the algorithm behaviour in the presence of degraded characters.

3. Detection and Classification

Trying to develop a method similar to the one of [4], but able to deal with characters cut on the left, down, right and up sides, neural networks (Multi-Layer Perceptrons) were built, one trained on a set of normal characters and an other trained on cut characters. But it soon appeared that a general criterion could not be determined to define a threshold for recognition rates defining which characters could be considered as cut characters. The reason is that, even if the relative magnitudes outputs are reliable enough (the class with the highest output comes first, followed by the class with the second highest output, ...), their absolute magnitudes are not properly ranked (the confidence level of the best output is often significantly less than 1). Detection of cut characters is therefore a problem on its own. Figure 2 shows the decomposition of the problem into two parts: the detection part first determines whether the character is processed "normally" or processed through a dedicated process for cut characters. Our efforts were mainly focused on the classification aspect but preliminary results about detection are also shown in the following section.

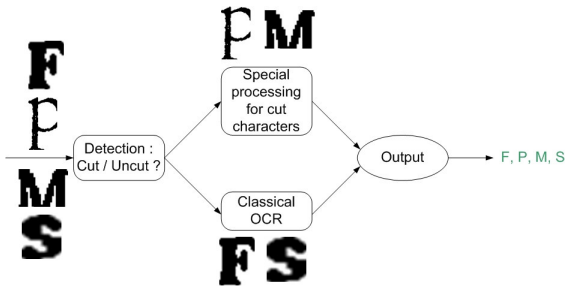

Figure 2. Decomposition of the whole problem into two sub-problems : detection and classification.

4. Detection

A basic algorithm was developed for cut characters detection, based on the analysis of the full picture containing the character to be processed. Characters are defined as cut characters if they are adjacent to the edges of the document, and the side of the cut assumed to be the edge of the picture adjacent to the character. Results seem to be encouraging, but some errors still occur, mainly for characters adjacent to more than one edge of the picture. However, simple post-processing block should improve the results.

This method is not useful in the case of physically degraded characters (ancient documents, coffee stains, ...), where the cuts can be anywhere in the picture.

5. Classification

Once the detection has been performed, the characters detected as uncut should be processed by typical algorithms (see [5]) and the detected cut characters need a special process.

Following the detection phase, all the cut sides of the (detected) cut characters are supposed to be known. Dedicated experts can therefore be used to classify those characters, as shown in Figure 3, where the four networks have been trained on characters cut on the same side (all the networks of this paper use contouring information, see Figure 4).

The situation is very different from how people deal with cut characters since people attempt to prolong the cut character mentally and compare it with a known character. Prolongation of cut characters should therefore help recognition and an algorithm based on this assumption was developed.

First trials for prolonging characters show difficulties to find the correct prolongation to apply, especially in the case of a noisy environment. Moreover, for a particular cut character, there may be several good prolongations, as shown

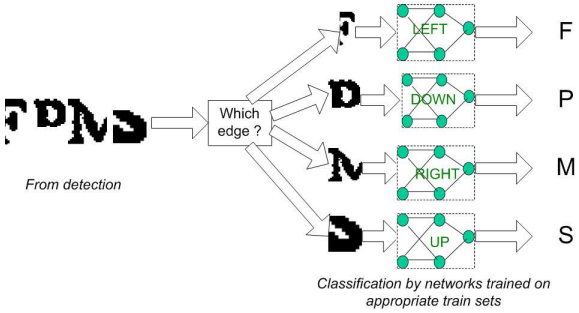

Figure 3. A classifier built on dedicated experts.

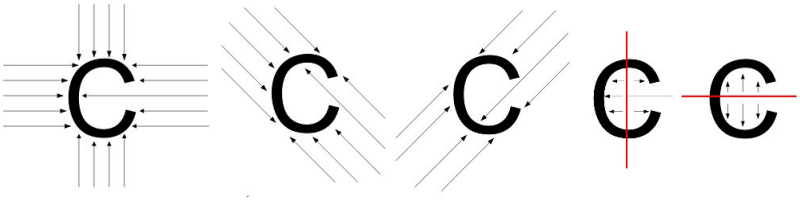

Figure 4. Use of different probes to detect the contouring information.

in Figure 5. Our novel algorithm developed acts differently, performing all possible prolongations and then selecting the best results.

Figure 5. Possible prolongations of a cut character (on the left). This character is more probably a "5", but it could be a "S", a "6", or a "G".

Figure 6 illustrates the algorithm. The (detected) cut characters follow two different channels. The first one is the one presented in Figure 3. The second one uses prolongation. Pending on the side of cut, the number of cuts and the number of connected components, several prolongations are performed. Figure 7 presents the different steps of a prolongation sequence. The skeletons of those prolongations are then extracted, in order to avoid variations due to inaccurate prolongations, as shown in Figure 8. Every skeleton of the prolonged characters is then sent to a network trained on uncut characters skeletons.

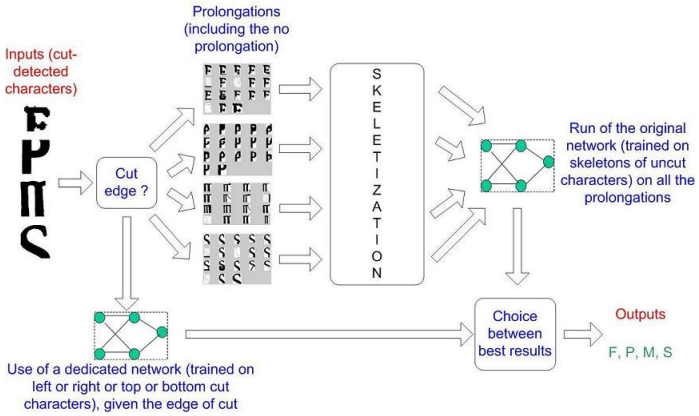

Figure 6. A classifier using prolongation

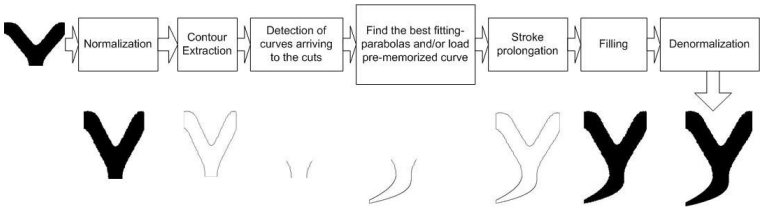

Figure 7. The different steps of a prolongation process

.

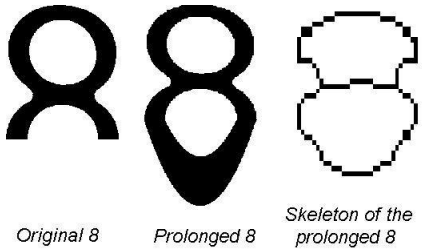

Figure 8. The skeletization process helps correcting imperfect prolongations.

Each prolonged character produces as many outputs as possible classes. The problem is now to choose the correct class amongst the best outputs. A selection amongst all the remaining character results is then performed. Because a selection based on the amplitude values of the network output is not reliable,

the selection uses relative magnitudes, in order to decrease the error rates: for each prolonged character, the best class is granted 4 points, the second class 2 points and the third 1 point. The candidates from the set of prolonged characters are then commingled with the candidates of the dedicated network to generate the final outputs.

6. Results

The database used for testing our algorithm contains characters from pictures taken with a poor-resolution digital camera. It contains about 500 cut characters, cut in the four sides.

Regarding the algorithm of Figure 3, the recognition rates reach 62% on the test database, and the use of prolonged character (algorithm of Figure 6) improves the recognition rates by 5%.

7. Conclusions and Future Work

First, prolongation seems to ameliorate results slightly. However, future work should be performed to improve recognition rates especially in the presence of a background noise. On the other hand, a post-processing block, using for example Viterbi algorithm and contextual information, should correct some current interpretation mistakes (like the mismatch between "5" and "S", . . .). The process of cut words should also be improved: for instance, a word cut on the left has missing letters and analysing letter sequences from the beginning of the word is irrelevant.

8. Acknowledgments

We want to thank the people of the TCTS, who provided us helpful advices.

References

[1] S. Ferreira, C. Thillou, B. Gosselin *From Picture To Speech: an Innovative Application for Embedded Environment*, Proc. of the 14th ProRISC workshop on Circuits, Systems and Signal Processing (ProRISC 2003), Veldhoven, Netherland, 2003.

[2] K. Fukushima, *Restoring partly occluded patterns: a neural network model with backward paths*, ICANN 2003, pp 393-400, 2003.

[3] B. Gosselin, *Application de reseaux de neurones artificiels q la reconnaissance automatique de caractere manuscrits*, PhD Thesis, Faculte Polytechnique de Mons, 1996.

[4] J.A.G. Nijhuis, A. Broersma, L. Spaanenburg, *A modular neural network classifier for the recognition of occluded characters in automatic license plate reading*, Proceedings FLINS2002, pp 363-372, Ghent, Belgium, 2002.

[5] C.Thillou, *Degraded Character Recognition*, DEA Report, Faculte Polytechnique de Mons, 2004.