Physical Layout Analysis of Complex Structured Arabic Documents Using Artificial Neural Nets

Karim Hadjar and Rolf Ingold

DIUF, University of Fribourg,
Chemin du Musée 3, 1700 Fribourg, Switzerland
firstname.lastname@unifr.ch

Abstract. This paper describes *PLANET*, a recognition method to be applied on Arabic documents with complex structures allowing incremental learning in an interactive environment. The classification is driven by artificial neural nets each one being specialized in a document model. The first prototype of *PLANET* has been tested on five different phases of newspaper image analysis: thread recognition, frame recognition, image text separation, text line recognition and line merging into blocks. The learning capability has been tested on line merging into blocks. Some promising experimental results are reported.

1 Introduction

In the field of document recognition many improvements have been made during the last decade. However, despite three hundred millions people around the world using Arabic language daily, there is still a lack especially in recognizing complex structured Arabic documents, such as newspapers or magazines. The major difficulty of such kind of documents is the variability of layout between newspapers and even different issues of the same newspaper.

The first methods of layout analysis for documents using Latin language focused on document structures [6, 10]. Recent works show a great interest in complex layout analysis [1, 2, 3] and also recommend the use of learning-based algorithms [5].

Currently known approaches rely on document models, which are either set up by hand or generated automatically in a previous learning step that needs a lot of ground-truthed data [8]. The drawback is that such models do not accommodate easily to new situations, a very common condition when dealing with complex document structures due to the variability of layout.

The Arabic language is known to be a difficult language for character recognition because of its specific features such as: Arabic alphabet is much richer than the Latin one, the form of the letter changes depending on its position inside the word, the words are written from right to left.

Therefore we tried to test the performance of the well known algorithms of segmentation and adapted them in order to treat complex structured Arabic documents newspapers [4]. Despite the encouraging results obtained, we believe that interactive

S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 170–178, 2004.
© Springer-Verlag Berlin Heidelberg 2004

incremental learning is an important issue in this context. It is one of the main goals of the CIDRE[1] project, which aims at building a semi automatic document recognition system that constructs its knowledge incrementally through interactions with the user.

In this paper we introduce *PLANET*, which stands for Physical Layout Analysis of Arabic documents using artificial neural NETs. It is an interactive recognition method for complex structured Arabic documents based on incremental learning.

This paper is organized as follows: in section 2 we present the principles of *PLANET*. Section 3 is devoted to the experimental part and the results obtained. Finally, in section 4 we conclude our work and give some perspectives for future work.

2 *PLANET* Principles

Up to now there is a shortage of tools either for the physical layout or for the logical layout extraction. These tools will help us building ground-truthed repositories useful for the document community. Nevertheless building them to behave in automatic way is not the right trend. First, the documents are getting more and more complex (different layouts, more colors, more typography …). Second it is hard to model a tool resolving all the problems of segmentation. It is more appropriate to build a tool treating all the phases of the physical layout and allowing incremental learning.

For these purposes *PLANET* has been constructed to allow:

- physical layout extraction (thread extraction, frame extraction, image text separation, text line recognition and line merging into blocks)
- improvement of the recognition ratio through an interactive incremental learning phase.

PLANET is composed of four sequential phases as illustrated in figure 1.

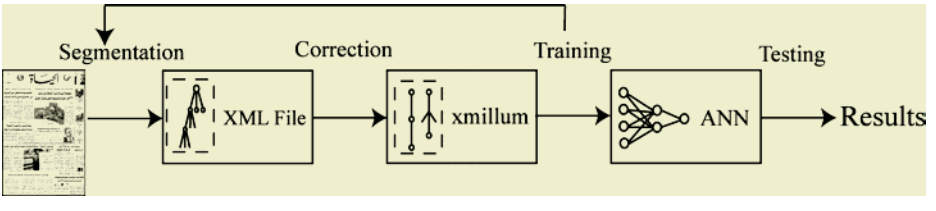

Fig. 1. Architecture of *PLANET*.

In the following subsections we review the four phases.

2.1 The Segmentation Phase

Our segmentation phase performs the following steps: thread extraction, frame extraction, image text separation, text line extraction and line merging into blocks.

[1] CIDRE stands for Cooperative and Interactive Document Reverse Engineering and is supported by the Swiss National Fund for Scientific Research, code 2000-059356.99-1

Threads are an essential part of the layout structure of newspaper; they serve as separators between columns of text or between different articles. Frames are special kind of paragraphs; they are paragraphs surrounded by rectangles. Our algorithm is modeled as a set of tools that can be used separately. A bottom-up approach based on connected components is used for image, thread and frames extraction. Connected components are also used for text line extraction after the use of RLSA. The line merging into blocks is done according to rules. More details about the thread extraction, frame extraction, image text separation, text line extraction and line merging into blocks are presented in [4]. The output of our segmentation algorithm is an XML file which describes the segmentation results concerning different components such as threads, images, frames, text lines extraction and line merging into blocks. A sample of the XML segmentation output is illustrated in figure 2.

```
<?xml version="1.0" encoding="UTF-8"?>
<segmentation image="Annahar_15_11_2003.tif">
 <Threads>
  <Thread x="827" y="955" w="3054" h="3" />
    ….
 </Threads>
 <Images>
  <Image x="361" y="1054" w="419" h="318" />
    ….
 </Images>
 <Texts>
  <Text x="413" y="15" w="8" h="1" />
   ….
 </Texts>
 <Frames>
  <Frame x="1010" y="4114" w="627" h="1082" />
 </Frames>
<Blocks>
  <Block x="1010" y="4114" w="627" h="1082" />
   ….
</Blocks>
 </segmentation>
```

Fig. 2. A sample of the XML segmentation output.

2.2 The Correction Phase

The correction phase, which follows segmentation is accomplished through an interactive process where users are able to interactively correct segmentation errors generated in the previous phase. We consider either over-segmentation or under-segmentation errors. For this purpose we use xmillum, which is our framework for cooperative and interactive analysis of document, which allows to visualize and to edit document recognition results expressed in any XML language [7]. User actions consist of merging or splitting in both directions, horizontally and vertically. This phase allows the creation of the training set needed by the artificial neural nets in the next phase. Figure 3 shows a screenshot of the correction with xmillum.

Fig. 3. Correction of the segmentation errors.

2.3 The Training Phase

The training phase follows the correction phase; at this level the training set is build for the artificial neural nets. In fact, we construct a training set for the initial document model. The initial document model is composed of samples of the correctly segmented blocks and samples of the user corrected blocks of each newspaper. Then, the artificial neural net is trained with the obtained training set. After that, the trained artificial neural net is copied three times, one for each newspaper, and trained with the corresponding newspaper training set. Figure 4 illustrates all the steps of the training phase.

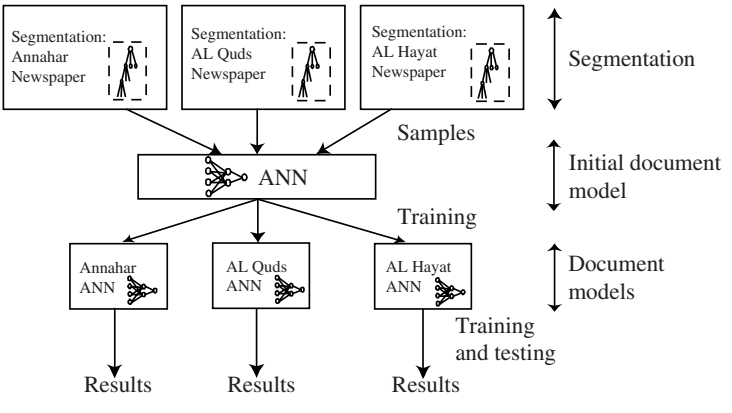

Fig. 4. The training phase.

2.4 The Test Phase

The test phase follows the training phase; at this level the three artificial neural nets are tested and their performances are evaluated. After segmenting each sample of each newspaper, a test file is constructed. This latter is composed of patterns containing features. The choice of these features used inside the artificial neural nets is described in the following subsection. Each artificial neural net generates an output file containing the computed output for each pattern. This output file is visualized by xmillum.

The neural net simulator used is Java Neural Nets Simulator (JavaNNS) [9]. Figure 5 shows an overview of our artificial neural net under JavaNNS.

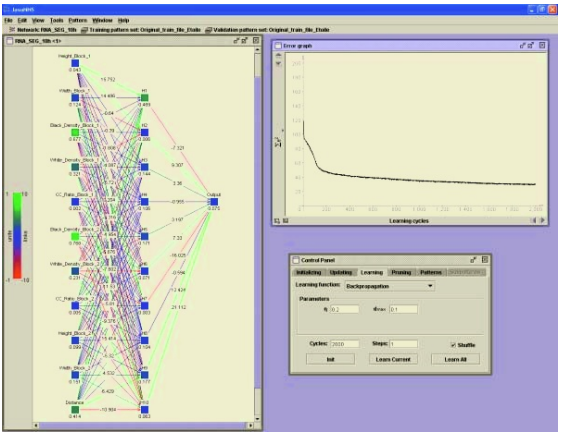

Fig. 5. Our artificial neural net under JavaNNS.

2.4.1 The Choice of Features

The choice of features is an important step; it leads generally to a good recognition ratio. At the beginning, our first goal was to improve the recognition ratio of merging lines into blocks obtained at the segmentation phase. The output of the line merging into blocks is a set of blocks containing errors due to either over-segmentation or under-segmentation. In order to correct such errors, we introduce to the artificial neural net this set by pair of blocks. Each block has the following features:

- width, height, black pixel density, white pixel density and connected component ratio,
- the distance between both blocks.

Our artificial neural net, a Multilayer Perceptron (MLP) [11], is connected in a feed-forward way and it is composed of three layers: input, hidden and output layer. The input layer is composed of 11 neurons; the first five neurons correspond to the features of the first block whereas the next five are those of the second block and the last neuron is the distance between the blocks. In order to determine the number of neurons of the hidden layer after multiple tests, we find that 10 neurons is the best configuration. Finally the output neuron is composed of 1 neuron stating whether to merge or to keep the blocks. A threshold is used inside the output layer of each artificial neural net in order to separate between the two output classes: merge and keep.

3 Tests and Results

Our first *PLANET* prototype has been implemented in Java with the integration of the neural net simulator JavaNNS. The segmentation and the features extraction took less

than thirty seconds whereas the learning phase inside JavaNNS is less than two minutes on a Pentium 4.

It has been tested on a set of pages from three Arabic newspapers: Annahar, AL Hayat and AL Quds. Figure 6 illustrates a page sample of AL Quds Arabic newspaper.


القدس العربي

مظاهرات حاشدة تؤيد طلب السيستاني الانتخابات

Fig. 6. Page sample of AL Quds newspaper.

The evaluation of *PLANET* has been performed on 90 pages of Annahar, AL Hayat and AL Quds newspaper. Table 1 shows the result of recognition for the AL Quds newspaper before the training. The corresponding results for Annahar and AL Hayat newspapers are illustrated in [4].

Table 1. AL Quds recognition results.

%	Thread	Frame	Image	Text line	Line merging into blocks
AL Quds	95,551	95,158	94,092	92,869	91,765

The low recognition rate for the text line is essentially due to a recurrent error which causes an ambiguity especially when diacritics of the first line and those of the second line are near to each other or merged. Since, the recognition rate of line merging into blocks is based on the results of the text recognition; the errors of the text line recognition are propagated to the next processing steps.

3.1 *PLANET* Training Phase

The training set for the initial document model is composed of samples of the well segmented blocks and samples of the user corrected blocks of three newspapers (5 pages from each one). After training the initial document model and duplicating it, one for each newspaper, we train again each artificial neural net with the corresponding newspaper training set. The number of pages used in the training phase for each newspaper is 25 pages for Annahar, 10 pages for AL Hayat and 10 pages for AL Quds. Figure 7 shows the *PLANET* recognition rate vs. the number of manipulations done by the user for the three newspapers.

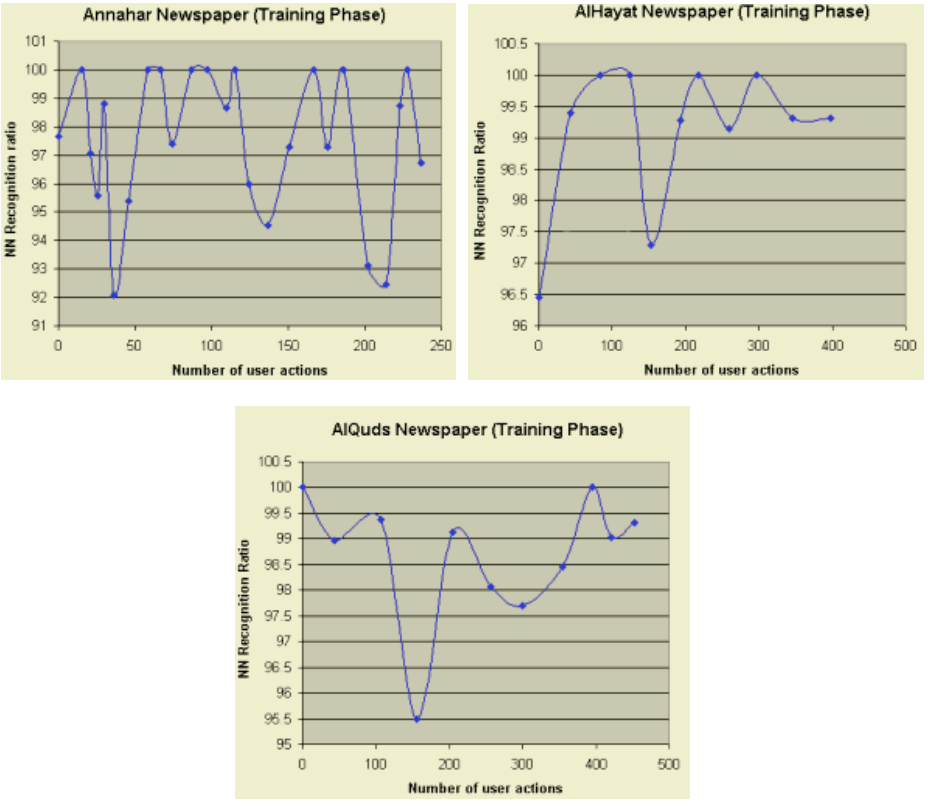

Fig. 7. *PLANET* recognition rate vs. the number of manipulation done by the user for the Annahar, AL Hayat and AL Quds newspapers.

The figures illustrated above are obtained this way: for each sample of one newspaper we test the recognition rate and we correct the misclassified patterns. Then we introduce these corrected patterns to the training set and we train again the artificial neural net.

From our observation we notice that there is a big variation of the layout for Annahar newspaper which corresponds to the variation of the recognition rate. This variation is less important for AL Hayat newspaper whereas for AL Quds is stable at the end. Table 2 shows *PLANET* average recognition rate for the three newspapers in the training phase.

Table 2. *PLANET* average recognition rate obtained in the training phase.

%	Annahar	AL Hayat	AL Quds
Average recognition rate	97,549	99,108	98,682

3.2 *PLANET* Test Phase

In the test phase we measure the performance of *PLANET*. A test set for each newspaper is constructed and tested within *PLANET*. First of all we have tested the initial document model with the training and test set. Table 3 shows *PLANET* average recognition rate obtained for the initial document model with the training and test set.

Table 3. *PLANET* average recognition rate for the initial document model.

%	Annahar	AL Hayat	AL Quds
Average recognition rate	96,667	96,857	96,572

Then we tested for each document model the recognition ratio with its corresponding test set. Table 4 shows *PLANET* average recognition rate with the test set.

Table 4. *PLANET* average recognition rate in the test phase.

%	Annahar	AL Hayat	AL Quds
Average recognition rate	97,963	98,343	99,051

First when comparing the results obtained in Table 3 and Table 4, we notice an improvement of the recognition rate for the line merging into blocks with the introduction of the learning in each document model. The following table illustrates the recognition rate obtained before and after the learning phase and the ratio of the improvement.

Table 5. *PLANET* results of recognition for Annahar, AL Hayat and AL Quds.

%	Annahar	AL Hayat	AL Quds
Segmentation only	95,217	91,438	91,765
Segmentation with learning	97,963	98,343	99,051
Ratio of improvement	2,348	5,167	8,677

We have also tested the cross recognition ratio: for example test the recognition ratio of Annahar with both the training sets of AL Hayat and AL Quds and vice-versa. Table 6 illustrates all the possibilities of the cross recognition ratio.

Table 6. *PLANET* cross recognition ratio.

Training set / Test set	Annahar	AL Hayat	AL Quds
Annahar	**97,963**	98,163	97,217
AL Hayat	95,630	**98,343**	96,336
AL Quds	96,632	98,081	**99,051**

The analysis shows that the Annahar, AL Hayat and AL Quds models are more specialized since the recognition ratio is decreasing when the training and the test set are not from the same newspaper. In fact each model has learned its own specific features.

4 Conclusion

In this paper we describe a document layout analysis method featuring interactive incremental learning named *PLANET*. Encouraging experimental results are reported concerning its application for recognizing complex structured Arabic documents.

The classification is driven by artificial neural net specialized in a document model. The first prototype of *PLANET* has been tested on five different phases of newspaper image analysis: thread recognition, frame recognition, image text separation, text line recognition and line merging into blocks. The learning capability has been tested on line merging into blocks.

We believe that *PLANET* can be used successfully as a tool to build ground-truthed repositories: users can, through some mouse clicks, easily correct segmentation errors and produce ground-truthed datasets.

Our future work on *PLANET* will focus the improvement of the model. We would like to test PLANET with other types of documents for example those in the Latin language, and may be other applications.

References

1. A. Antonacopoulos, B. Gatos and D. Karatzas. “ICDAR 2003 Page Segmentation Contest”. Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, August 2003, pp. 688-692
2. B.Gatos, S.L. Mantzaris and A. Antonacopoulos. “First international newspaper segmentation contest”. Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, September 2001, pp. 1190-1194
3. K. Hadjar, O. Hitz and R. Ingold. “Newspaper Page Decomposition using a Split and Merge Approach”. Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, September 2001, pp. 1186-1189
4. K. Hadjar and R. Ingold, “Arabic Newspaper Page Segmentation”, Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, August 2003, pp. 895-899
5. K. Hadjar, O. Hitz, L. Robadey and R. Ingold, “Configuration REecognition Model for Complex Reverse Engineering Methods: 2(CREM)”, Proceedings of the 5th International Workshop on Document Analysis Systems, DAS2002, Princeton, USA, August 2002, pp. 469-479
6. R.M. Haralick. “Document image understanding: Geometric and logical layout”. Proceedings Internet. Conf. On Computer Vision and Pattern Recognition, 1994, pp. 385-390
7. O. Hitz, L.Robadey, and R. Ingold. “An architecture for editing documents recognition results using xml technology”. Proceedings of the 4th International Workshop on Document Analysis Systems, DAS2000, Rio de Janeiro (Brazil), December 2000, pp. 385-396
8. J. Hu, R. Kashi, D. Lopresti, G. Nagy and G. Wilfong. “Why table ground truthing is hard”. Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, September 2001, pp. 129-133
9. JavaNNS. http://www-ra.informatik.uni-tuebingen.de/software/JavaNNS/welcome_e.html
10. G. Nagy. “Twenty Years of Document Image Analysis in PAMI”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No 1: January 2000, pp. 38-62
11. P.D. Wasserman. “Neural Computing: Theory and Practice”, Van Nostrand Reinhold, 1989