2011 International Conference on Document Analysis and Recognition

Minimizing user annotations in the generation of layout ground-truthed data

Karim Hadjar
Department of Multimedia
Ahlia University
Manama, Kingdom of Bahrain
karimh@ahliauniversity.edu.bh

Rolf Ingold
Department of Informatics
University of Fribourg
Fribourg, Switzerland
rolf.ingold@unifr.ch

***Abstract*—**This paper describes the adaptation of a previously developed document recognition framework called PLANET (Physical Layout Analysis of complex structured Arabic documents using artificial neural NETs) into a groundtruthing system for complex Arabic document images [8]. PLANET is a layout analysis tool for Arabic documents with complex structures allowing incremental learning in an interactive environment. Artificial neural nets drive the classification of homogeneous text blocks. We have observed that when users use PLANET for groundtruthing, the number of interactive corrections is quite large. In order to reduce user intervention and to make use of PLANET as a groundtruthing system we have adapted its architecture.

Keywords: Ground Truth, Physical Layout Extraction, Datasets, Document Image, Artificial Neural Networks, Arabic Newspapers

I. INTRODUCTION

In the field of document recognition many improvements have been made during the last decade. However, despite the availability of many public document image datasets, from page segmentation to OCR, with ground truth information we have noticed few are shared and most of them have a proprietary format. Another issue is that in order to compare the performance of one class of documents with two datasets a time consuming process is required, and also some of the datasets are relative to a specific type of class of documents.

The creation phase of datasets is costly and requires a lot of time and this is why few ground-truthed datasets are available. Two directions should be followed by the document researchers' community in order to establish a valuable ground-truthed datasets; the first one is to agree on an open format, like XML for example, and the second one is to make sure that ground-truthed datasets will be adaptive to any document class.

Creating a ground-truthed tool which behaves in an automatic way is not the right way to take and this is why PLANET was modeled as a semi-automatic process. PLANET is a system that allows the extraction of the physical structure for Arabic complex documents allowing incremental learning in an interactive environment. In PLANET, the user will interactively through some mouse clicks easily correct segmentation errors and produce ground-truthed datasets.

Inside PLANET, there are essentially two steps: the first one is to run the recognition engine while the second is that the user will correct segmentation errors which allow incremental learning. After completing these steps a ground-truth is produced. Even so the incremental learning based on Artificial Neural Networks drastically reduces the number of corrections made by the user and we do believe that we can reduce this number more by reversing the process inside PLANET.

In this paper, we describe how we improved the performance of PLANET by reversing the steps which have allowed the creation of ground-truthed datasets in a much easier way than before.

This paper is organized as follows: in section 2 a short overview of ground-truthed datasets and related work is described. Section 3 presents the details about the architecture of PLANET. In section 4, we deeply discuss the minimization of user annotations' process. The results of our experiments are discussed in Section 5 and Section 6 concludes this paper.

II. RELATED WORK

Since 2001, ICDAR has encouraged the document analysis community to register for the document recognition contest. In fact, small datasets were created for that purpose [1, 2, 3, 4]. We have participated in that contest in 2001 [10], which had few participants registered. Since that year the number of participants increased significantly. We remember, in ICDAR'01, the participants of the contest described the difficulty of creating such datasets.

Historical documents have become an entire field of research in the document community and some attempts have been made towards the creation of ground truth. Fischer et al. [6] discussed the creation of ground truth for handwriting recognition in historical documents. It is a semi-automatic creation mode of ground truth applied to old manuscripts that takes into consideration noise and transcription alignment.

Strecker et al. [13] presented an approach that allows the creation of ground truths datasets. Synthetic images are generated from personalized newspapers. In order to make the system much more robust, the synthetic images are degenerated and scanned then an alignment technique is used in order to align the ground truth. The automatic generation of the layout of the personalized newspapers is described in detail in [14].

Regarding page segmentation Heroux et al. [11] developed a system allowing the generation of synthetic images with their corresponding ground truth. In [9] we have

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.147

shown the importance of a semi-automatic process for the creation of ground truth by including incremental learning using Artificial Neural Networks. After that we tried to adapt PLANET in order to handle logical labeling: LUNET [7]. These latter can both be used in the creation of ground truths.

Doermann et al. [5] developed a document image annotation tool titled GEDI which allows the creation of ground truths by annotating a document image using zones. A set of specific tools are provided within the user interface in order to annotate the different zones. The output of GEDI is an XML file containing the annotation made by the user. GEDI in overall is an annotation tool, even that optionally can generate connected components and Run Length Encoding.

Our approach allows the creation of ground truthed datasets using a semi automatic process that includes incremental learning in order to reduce the intervention of the user in the creation process.

III. PLANET PRINCIPLES

As stated in section 1, PLANET is a system that allows the extraction of the physical structure for complex Arabic documents allowing incremental learning in an interactive environment. It includes many models that integrate learning in our document analysis system. In fact, we have dedicated models to a class of documents and their tasks are to learn the features of this class. Besides these dedicated models, we have a universal model dedicated to an abstract class, including all documents. The architecture of PLANET is illustrated in figure 1.

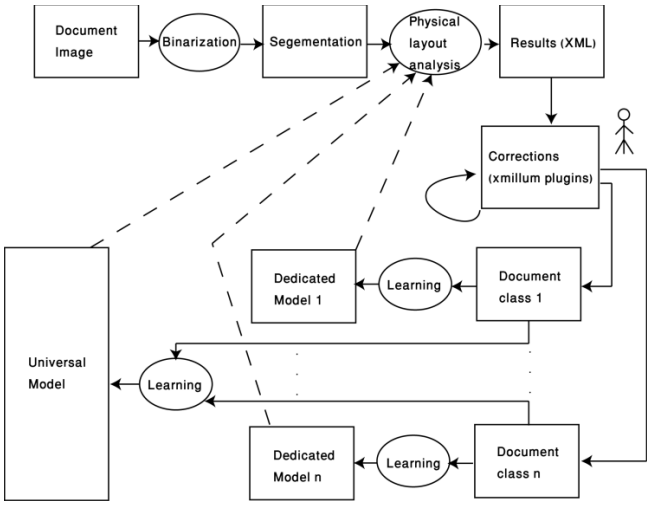

Figure 1. The architecture of PLANET.

PLANET enhances the document analysis system by allowing the creation of ground truths datasets. Due to the variability between document classes, PLANET is constructed so that it is capable of treating many classes of documents. Thus, for each class of documents, a dedicated model assigned to it. This dedicated model will learn the features of the document class and will adapt itself when a change occurs.

Each dedicated model receives as input the corrections made by the user. These corrections are included in the learning set of each dedicated model.

Once the learning has been done, the dedicated model becomes specialized in the class of documents that it belongs to. Currently, there are new classes of documents that appear: newspapers, magazines… With dedicated models we ensure PLANET*'s* extensibility. If we would like to add a new class of documents we assign a dedicated model to it. Nevertheless, this added dedicated model doesn't have a memory of the learned corrections made by the user. We have created the universal model to ensure that each new added dedicated model to PLANET possesses a memory.

The universal model is initially created from a set of document classes. From each class we take samples that are issued from the corrections, and then we train the universal model. Once the learning of the universal model is done, we reflect this knowledge to each dedicated model.

The models defined in PLANET namely the universal model and the dedicated models are composed of artificial neural networks. These latter are trained on the basis of samples issued from the corrections made by the user on segmentation results.

The artificial neural network of the universal model is trained first. Once the training of the universal model is completed with success, we transfer the universal knowledge to each artificial neural network of each dedicated model. The transferred knowledge is indeed the training set of the universal model, and it allows the artificial neural networks of the dedicated models to possess an initial knowledge.

The correction of layout errors is accomplished through an interactive process where users are able to interactively correct them through xmillum plugins. Xmillum [8] is our framework for cooperative and interactive analysis of document, which allows visualizing and editing document recognition results expressed in any XML language. Once all the errors are corrected, an XML file is generated and it comprises all the actions done by the user.

Experimental results that we have obtained with PLANET are encouraging; nevertheless we have noticed that the number of corrections that the user has to do is quite large. This will discourage users from using PLANET as a groundtruthed dataset system and that is why we have revised PLANET's architecture.

IV. MINIMAZATION OF USER ANNOTATION' PROCESS

A groundtruthing dataset system should be semi-automatic, interactive and adaptive. All these latter are bundled within PLANET. But we have noticed that we can still improve the performance of PLANET not in terms of recognition ratio but in terms of efficiency of user actions by reducing them. In order to do this we have reversed some of the modules of PLANET.

Actually PLANET starts by applying the physical layout analysis which is enriched with the learning capabilities of

the different models then there is the intervention of the user which corrects the physical layout errors. We recommend switching these two modules. The first one is the interaction phase in which the user annotates the document image through xmillum plugins that we have adapted. The second one is the learning phase. In this module there is the extraction of knowledge from the abstraction of the annotation made by the user which feeds both the dedicated model and the universal model. PLANET-2 is illustrated in figure 2.

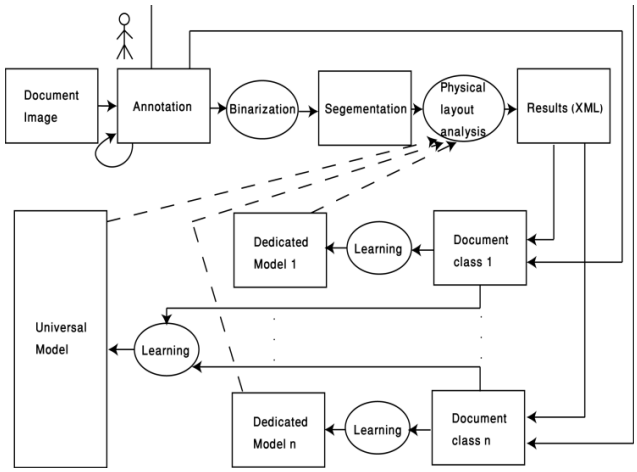

Figure 2. PLANET-2.

In the annotation phase the user defines the blocks of the document image using xmillum plugins. A block is defined with mouse clicks over the document image after selecting the type of tools: box, or polygon, or ellipse. In case of manipulation errors the user can undo the action or choose a split and merge tool from the toolbox to correct the zone or the block. The result of the annotation phase is an XML file which describes the actions of the user on the document image. A sample of the XML annotation output is illustrated in figure 3.

After completing the annotation phase, our segmentation algorithm is executed. The output of this latter 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 XML segmentation output is illustrated in figure 4.

```
<?xml version="1.0" encoding="UTF-8"?>
<segmentation image="Annahar_15_11_2010.tif">
  <Blocks>
    <Block x="200" y="300" w="627" h="235"/>
    <Block x="230" y="336" w="500" h="321"/>
    <Block x="340" y="546" w="326" h="800"/>
    <Block x="815" y="1250" w="452" h="760"/>
    <Block x="900" y="1800" w="642" h="1082"/>
    ...
  </Blocks>
</segmentation>
```

Figure 3. A sample of the XML annotation output.

All the models present in PLANET are Multilayer Perceptrons (MLPs) [8] which are connected in a feed-forward way. The artificial neural network of the universal model is trained first. Once the training of the universal model is completed with success, we transfer the universal knowledge to each artificial neural network of each dedicated model. The transferred knowledge is indeed the training set of the universal model, and it allows the artificial neural networks of the dedicated models to possess an initial knowledge.

```
<?xml version="1.0" encoding="UTF-8"?>
<segmentation image="Annahar_15_11_2010.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>
```

Figure 4. A sample of the XML segmentation output.

The choice of features is an important step; it leads generally to a good recognition ratio. The features, that we chose form the annotation phase done by the user, are relative to the physical attributes of the blocks. We have improved the features compared to the one that we have chosen in our previous work [8]. In fact, we extracted the height, width, width/height ratio, black pixel density, white pixel density and connected component ratio from the block. Since the block is represented by a rectangle, the first two features represent the height and the width of this rectangle. The third feature represents the width over the height ratio of the rectangle. The forth and the fifth features respectively represent the number of black and white pixels inside the rectangle. Finally, the sixth feature represents the number of connected components included inside the rectangle. All these extracted features will be used for the input layer of the artificial neural networks of both the universal model and the dedicated models.

V. EXPERIMENTS AND RESULTS

The evaluation of PLANET-2 has been performed on 60 pages from two document newspaper classes. The distribution of these pages into each document newspaper class is as follows:

- 30 pages of document class AL Quds are used. Figure 5, illustrates a page sample from this newspaper.
- 30 pages of document class Annahar are used. Figure 6, illustrates a page sample from this newspaper.

Figure 5. Page sample of AL QUDS newspaper.

Figure 6. Page sample of ANNAHAR newspaper.

In order to compare the number of actions done by the user in PLANET and in PLANET-2 we have run the two versions. Table 1 shows the number of user actions for the version with two document newspaper classes.

TABLE I. NUMBER OF ACTIONS IN PLANET AND IN PLANET-2

Number of user actions for the two document classes	*PLANET*	*PLANET-2*
AL QUDS	778	619
ANNAHAR	767	608

From our observation we notice that with PLANET-2 the number of user action for the set of pages is less than with PLANET. The reduction of the user actions with PLANET-2 is significant; we have reached an average of 25% of reduction. The number of user action in PLANET is larger than in PLANET-2 because in PLANET the user will apply corrections by splitting or merging in case we have over segmentation. We have also noticed that the difference of user action at the beginning is big with PLANET while at the end of our test, it is small thanks to the learning capabilities of PLANET.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we describe the modification of PLANET (Physical Layout Analysis of complex structured Arabic documents using artificial neural NETs) into a groundtruthing system for complex Arabic document images.

We have restructured PLANET in order to make the creation of ground-truth datasets much easier. A significant improvement, by reducing the number of user actions has been achieved thanks to restructuring PLANET.

We believe that PLANET-2 can be used successfully as a tool to build ground-truthed datasets: users can, through some mouse clicks, produce ground-truthed datasets.

Our future work on PLANET-2 will focus on testing with other types of documents for example those in the Latin language, and may be other applications.

REFERENCES

[1] A. Antonacopoulos, D. Bridson and B. Gatos, "Page Segmentation Competition". In Proc. of the 9th International Conference on Document Analysis and Recognition (ICDAR'2007), Curitiba, State of Parana, Brazil, September 2007, pp. 1279-1283.

[2] A. Anronacopoulos, D. Bridson and B. Gatos, "Page Segementation Competition". In Proc. of the 8th International Conference on Document Analysis and Recognition (ICDAR'2005), Seoul, Korea, September 2005, pp. 75-79.

[3] A. Anronacopoulos, D. Bridson and B. Gatos, "Page Segementation Competition". In Proc. of the 7th International Conference on Document Analysis and Recognition (ICDAR'2003), Edinburgh, Scotland, UK, August 2003, pp. 688-692.

[4] B. Gatos S. Mantzaris and A. Anronacopoulos, "First International newspaper segmentation contest". In Proc. of the 6th International Conference on Document Analysis and Recognition (ICDAR'2001), Seattle, WA, USA, September 2001, pp. 1190-1195.

[5] D. Doermann, E. Zotkina and H. Li, "GEDI – A Groundtruthing Environment for Document Images". In Proc. of the 9th IAPR International Workshop on Document Analysis Systems (DAS'2010), Boston, USA, June 2010, pp. 519–522.

[6] A. Fischer, E. Indermuhle, and H. Bunke, "Ground Truth Creation for Handwriting Recognition in Historical Documents" In In Proc. of the 9th IAPR International Workshop on Document Analysis Systems (DAS'2010), Boston, USA, June 2010, pp. 3–10.

[7] K. Hadjar and R. Ingold, "Logical Labeling of Arabic Newspapers using Artificial Neural Nets**".** In Proc. of the 8th International Conference on Document Analysis and Recognition (ICDAR'2005), Seoul, Korea, September 2005, pp.426-430.

[8] K. Hadjar and R. Ingold, "Physical Layout Analysis of Complex Structured Arabic Documents using Artificial Neural Nets". In Proc. of the 6th IAPR International Workshop on Document Analysis Systems (DAS'04), Florence, Italy, September 2004, pp. 170-178.

[9] K. Hadjar and R. Ingold, "Arabic Newspaper Page Segmentation". In Proc. of the 7th International Conference on Document Analysis and Recognition (ICDAR'2003), Edinburgh, Scotland, UK, August 2003, pp. 895-899.

[10] K. Hadjar, O. Hitz and R. Ingold, "Newspaper Page Decomposition using Split and Merge Approach". In Proc. of the 6th International Conference on Document Analysis and Recognition (ICDAR'2001), Seattle, WA, USA, September 2001, pp. 1186-1189.

[11] P. Heroux, E. Barbu, S. Adam, and E. Trupin, "Automatic ground-truth generation for document image analysis and understanding". In Proc. of the 9th Int. Conf. on Document Analysis and Recognition (ICDAR'2007), Washington DC, USA, 2007, September 2007, pp. 476-480.

[12] G. Nagy, and D. Lopresti, "Interactive document processing and digital libraries". In In Proc. of the 2nd International Workshop on Document Image for Libraries (DIAL'2006), Lyon, France, April 2006, pp. 2-11.

[13] T. Strecker, J. van Beusekom, S. Albayrak, and T. Breuel, "Automated Ground Truth Data Generation for Newspaper Document Images". In Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR'2009), Barcelona, Spain, July 2009, pp. 1275-1279

[14] T. Strecker, and L. Hennig, "Automatic layouting of personalized newspaper pages". In Proc. of Int. Conf. on Operations Research, University of Augsburg, Germany, September 2008, pp. 469-474.