A Comparative Study On Optical Modeling Units For Off-line Arabic Text Recognition

Mohammed Faouzi BenZeghiba
A2iA S.A.
39 rue de la Bienfaisance
75008 - Paris - France
Email: mfb@a2ia.com

Abstract—The role of the optical model in a text recognition system is to model the textual information written in image documents. This paper compares the performance of four Arabic optical modeling units in a Multi-Dimensional Long Short-Term Memory based state-of-the-art Arabic text recognition system. These units are: 1) The isolated characters, 2) Extended isolated characters with the different shapes of Lam-Alef (Yi), 3) The character shapes within their contexts and, 4) The recently proposed sub-character units that allow sharing similar patterns in the different character shapes. Experiments are conducted on six tasks using Maurdor and Khatt databases. For a fair comparison, optical models are trained from scratch. The decoding is performed using 1) the predictions of the optical model only and, 2) combined with a 3-gram hybrid word/part-of-Arabic word language model. Results in terms of Word Error Rate show that the best results are generally obtained with systems using isolated characters as the basic modeling units, although differences in the performance among different systems are negligible.

I. INTRODUCTION

Automatic text recognition is the task of identifying and recognizing automatically words in written documents (i.e; scanned images) and convert them to editable text by machines. State-of-the-art text recognition systems integrate several components, including image preprocessing, feature extraction, optical model and language model. Although all of these components are important, this work focuses on the optical model.

The role of the optical model is to model and represents the relationship between basic optical units and the text in images. Several optical modeling techniques are investigated. For Arabic script, the most widely used -in recent years- are Hidden Markov Models (HMM) [1][2][3] and Artificial Neural Networks (ANN) [4][5][6]. The work on the optical model includes several topics. However, in this work, we are particularly interested in the choice of the set of basic optical units, which are the fundamental writing units used to represent any Arabic text.

The natural way to define the optical units is to use the set of different characters in the language. This choice is not always practical. Some languages, such as Chinese have a huge number of characters, which might arise some modeling issues [7]. A more natural criterion will be something based on the visual description of the written text. In this case, if enough training data for each word in the recognition vocabulary is available, then considering the whole word or the sub-word as a basic optical unit might be a good choice. In [4], Part-of-Arabic-Words (PAW) are used as optical units to train a Multi-Layer Perceptron (MLP) on the IFN/ENIT database. However, for large vocabulary applications (such as the recognition of historical documents), the use of words as optical units is not practical. In such applications, the vocabulary is dynamic, so new task-specific words need to be added to the vocabulary, which requires the search for adequate training data. A more practical approach is to use optical units that are smaller than words. To be effective, these units need to be easy to learn from training data and generalizable, so every word in the vocabulary can be derived from an inventory of these units.

For Arabic language, the use of the alphabet characters as the basic optical units might be seen as the most obvious and appropriate choice. However, since the shape of the character changes according to its position within the word, it might be more appropriate to use the shape of characters as the basic optical units, with the cost of increasing slightly their numbers. In [8], a comparison between these two types of units was conducted. It has been found that with sufficient amount of data, units based on the shape of characters perform better than characters in their isolated shapes. A compact set of units called sub-characters, derived from the character shapes was first proposed in [9] and enhanced in [10]. Sub-character units share common patterns in different character shapes. Experiments on the IFN/ENIT database have shown that systems using sub-character units outperformed those using character shapes.

It might be argued that the best set of units depends on several factors, in particular the framework of the optical model, the amount of training data and the nature of the task (isolated or continuous, printed or handwritten text recognition). Compared to previous similar works, the main contributions of this work are: First, in previous studies [8][10], optical models were based on the HMMs framework. These models try to capture the characteristics of each unit, making them more suitable for modeling the character shape. However, state-of-the-art Arabic text recognition systems are based on the recently proposed Multi-Dimensional LSTM Recurrent Neural Networks (MDLSTM-RNN) [11]. Since these models are discriminative (they try to model the frontier between units), and the prediction of the output units is conditioned on the whole input image, findings of the previous studies might not
be valid. Comparison of optical units within the MDLSTM-RNN framework has not been conducted yet. Second, in previous works, the task was generally limited to isolated word recognition, except for [10]. In this work, experiments are conducted on three benchmarking databases with continuous text lines. These databases contain more variabilities in terms of writing type, writing style and lexical content.

The paper is organized as follows: Section II describes the different optical units that have compared in this work. Section III describes the experimental setup, including databases, systems and the decoding process. Section IV reports the performance of the optical units and Section V concludes the work with description of possible extensions in future work.

II. ARABIC OPTICAL Modeling Units

This section describes different optical units that are compared in this work. Description will be limited to those characters that are specific to the Arabic language. Digits and special symbols (i.e; (, ), [ ] { }) that might accoutered in several written documents are not considered here, even if they are part of the final set of optical units.

A. Character units

Arabic alphabet consists of 28 basic characters and 8 additional characters. This makes up a total of 36 characters or optical units as shown in Table II.

<table>
<thead>
<tr>
<th>Character</th>
<th>Isolated</th>
<th>Beginning</th>
<th>Middle</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>س</td>
<td>س</td>
<td>ف</td>
<td>غ</td>
<td>ط</td>
</tr>
<tr>
<td>ف</td>
<td>ف</td>
<td>غ</td>
<td>د</td>
<td>ن</td>
</tr>
<tr>
<td>غ</td>
<td>غ</td>
<td>ط</td>
<td>د</td>
<td>ن</td>
</tr>
</tbody>
</table>

Using these units, the word """" (بدلاً من دوام آتى) """" will be transcribed as """"بدلاً من دوام آتى"""". More examples are shown in Table I. The first row shows four Arabic words written in cursive style. Their transcriptions in terms of isolated character units are given in the second row. This set has a relatively small number of units, making it more suitable for cases where the amount of training data is very limited, or to build compact recognizers, in particular with HMM models. Hereafter, this set will be simply referred to as Chars.

B. Extended character units

In the Chars, the different variants of the ligature Lam-Alef ( ﯋) with or without hamza are transcribed with two characters, the Lam followed by Alef ( ﯋) with or without hamza. Given the special shape of this ligature, it might be more appropriate to use a separate unit for each of its variants ( ﯋, ﯋ and ﯋), resulting in a set with 40 optical units. Using this set, the word """"بُنْيَان"""" will be transcribed as """"بِنَبَيْنَيْ"""" and not """"بُنَيْنَ"""". More examples are shown in Table I (second row). This set of units was used in the work described in [1]. Hereafter, this set of optical units will be referred to as E-Chars units.

C. Character shape units

Arabic script is cursive for both handwritten and printed text. Depending on its position within a word (i.e, isolated, beginning, middle or end), the same character can have up to four different shapes. Table III shows examples of some Arabic characters and their shapes. It can be seen that the character ﯋ has only two shapes, isolated (similar to beginning) and final (similar to middle) shapes.

<table>
<thead>
<tr>
<th>Character</th>
<th>Isolated</th>
<th>Beginning</th>
<th>Middle</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>س</td>
<td>س</td>
<td>ف</td>
<td>غ</td>
<td>ط</td>
</tr>
<tr>
<td>ف</td>
<td>ف</td>
<td>غ</td>
<td>د</td>
<td>ن</td>
</tr>
<tr>
<td>غ</td>
<td>غ</td>
<td>ط</td>
<td>د</td>
<td>ن</td>
</tr>
</tbody>
</table>

Using shapes of characters as optical units, the word """"(بدلاً من دوام آتى) """" will be transcribed as """"(بدلاً من دوام آتى)"""". More examples are shown in Table I (third row).

It might be argued that using these units, modelisation of the written text will be more precise and accurate. So, better performance can be expected, in particular when sufficient training samples for each character shape is available. However, as pointed in [8], some of these units might appear rarely in the training data, which might result in poor models for those units. A variant of this set was investigated in that same work, where similar character shapes are grouped together (according to their positions in the word) and modeled with the same unit. More precisely, beginning and middle shapes of the same character are modeled with the same unit, similarly for the end and isolated shape. Such grouping has not investigated in this work.

The number of different shapes of all Arabic characters is about 125 shapes, which is more than 65% higher compared to the size of Chars units. Hereafter, this set of optical units will be referred to as Shape units.
D. Sub-characters

To search for better trade-off between Char and Shape units, The authors in [9] and [10] have proposed the use of sub-character units. These units share similar patterns (or segments) between different shapes of the same character as well as between different characters. Each character shape is then split into two or more common patterns that are previously defined. These patterns called sub-characters. Examples showing how these sub-characters are used to represent character shapes are shown in Table IV. Table I shows some examples at the word level. This approach leads to a reduction (compared to Shape units) in the number of optical units. But this number is still higher compared to Char units. These units will be referred to as Sub – Char.

<table>
<thead>
<tr>
<th>Shape of character</th>
<th>Sub Characters</th>
<th>Shape of character</th>
<th>Sub Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>ف</td>
<td>ف</td>
<td>ق</td>
<td>ق</td>
</tr>
<tr>
<td>ف</td>
<td>ف</td>
<td>ق</td>
<td>ق</td>
</tr>
<tr>
<td>ق</td>
<td>ق</td>
<td>ق</td>
<td>ق</td>
</tr>
</tbody>
</table>

III. EXPERIMENTAL SETUP

Results of previous similar studies [8][10] are reported on the APTI database for the first work and on the IFN/ENIT and the printed part of the Khatt database for the second. The APTI and IFN/ENIT are more suited for isolated word recognition task, making the generalization of the findings to more complex task rather difficult. Moreover, the interaction between the optical model and the language model was not clearly investigated.

A. Database description

Efficiency of different optical units is evaluated on three benchmarking tasks: the Khatt handwritten, the Maurdor printed and the Maurdor handwritten tasks.

1) KHATT task: KHATT is a freely available offline handwritten text database [12]. It contains 2000 unique randomly selected paragraphs with different text contents, and 2000 fixed paragraphs with the same text content. Line text images with their annotations are provided. So line detection and segmentation was not required. For a fair comparison with the other published works, experiments are conducted using both random and fixed paragraphs. However, to make the task more harder and realistic, experiments using only random paragraphs are also conducted. Some statistics about the train and test data can be found in [6].

2) MAURDOR tasks: The Maurdor database [13] consists of multi-lingual text document images with several variant topics. Two tasks are performed on this database, The Maurdor printed and the Maurdor handwritten tasks. Annotations are available at the paragraph level. So, a line detection and segmentation step is required. Detail about this procedure can be found in [14]. Some statistics about the train and test data for each task can be found in [6]. For simplicity, the Maurdor handwritten Arabic and the Maurdor printed Arabic tasks will be referred to as M-HWR, and M-PRN, respectively.

B. System description

For comparison purposes, all systems use the same architecture of the MDLSTM-RNN based optical model. They differ only by the size of the output layer which is related to the number of the optical units.

1) Optical model: Optical models make use of a Multi-Directional LSTM Recurrent Neural Networks (MDLSTM-RNN) [11]. The architecture of this model in terms of layer types and their configurations is shown in Figure 1. The only difference between optical models within the same and between tasks is the size of the softmax layer, which corresponds to the number of optical units (which is task-dependent) augmented with the special blank unit.

The softmax layer is used to estimate the posterior probabilities of the optical units given the input image. Additional Connectionist Temporal Classification (CTC) layer is used during the training. The advantage of using the CTC, is that the annotation (the target outputs) has to be given only as a sequence of the corresponding optical units without further segmentation. For fair comparison, and to remove any bias due to some a priori information, all optical models are trained from scratch (i.e; no adaptation). To improve the generalization capability of the model, a dropout technique is used [15]. For each task, the development dataset is used as validation data to select the model with the best parameters.
Fig. 1. The architecture of the MDLSTM-RNN based optical model used in this work. This architecture is common for all tasks, except for the softmax layer which depends on the task and the size of the optical units set.

For the Khatt task and for each set of the optical units, two optical models are trained, one on the random paragraphs only, and the other on both random and fixed paragraphs.

Table V reports the size of the softmax layer for each task, and for each optical unit set. It can be seen that the use of Sub–Char reduces the number of optical units by almost 30% compared to the character shape units. This reduction is more important when the Char and E–Char units are used (more than 50%). From practical point of view, this reduction might appear interesting in terms of speed and memory usage, within the HMMs framework. However, within the MDLSTM-RNN framework, this reduction has almost no effect.

2) Language model: To make this comparison study more complete, different optical unit sets are evaluated within a complete state-of-the-art Arabic text recognition system. The goal is to evaluate the interaction between the optical model and the language model. For this study, a hybrid word/Part-of-Arabic-Word (PAW) LM is used [6]. This hybrid model integrates both the most frequent words, and PAWs, resulted from decomposing the less frequent words. This hybrid LM has shown to be superior to standard word LM, in particular for tasks with relatively high out-of-vocabulary rate. For a given task, the same 3-gram hybrid flat LM, generated on the task-specific data is used in all systems (i.e, independently on the type of optical units).

C. Decoding

The outputs of the optical model can be interpreted as posterior probabilities of optical units conditioned on the input image. These posterioriors are normalized by the optical unit priors and used as HMM state emission probabilities for decoding.

The decoding is performed through a Weighted Finite-State Transducers (WFST) [16] composed of several WFST representing different system components.

\[
D = H \circ C \circ L \circ G
\]

where H represents the Context-Independent Hidden Markov Model (CI-HMM) states, where each state is associated with one optical unit (i.e; output of the MDLSTM), C are the rules defining the context (no context is used in this work), L is the spelling lexicon that maps a sequence of optical units to a single word and G is the grammar (i.e; language model).

The modeling unit for the LM is a word or PAW written with isolated characters. The conversion from the type of optical units to isolated characters is done within the lexicon L, where each input (a sequence of optical units) is associated with an output representing the word written with isolated characters. So, the recognition hypothesis of all systems is a sequence of words written with isolated characters. Generation of the decoding graph and the decoding itself are performed with the Kaldi toolkit [17].

Performances are reported in terms of Word Error Rate (WER). The WER is computed from the alignment between
recognition the hypotheses and the annotations (ground truth):

\[ WER = \frac{S + D + I}{N} \]  

(2)

where \( S \), \( D \) and \( I \) are the number of substitutions, deletions and insertions, respectively. \( N \) is the total number of words in the reference.

IV. RESULTS

For a fair comparison, systems are evaluated under two different conditions. In the first condition, only outputs of optical models are used. The goal is to assess how good a set of basic units is, without any interaction with external/complementary information. In the second condition, optical model outputs are combined with language model scores, and using some optimized parameters such as acoustic scale factor and priors scale. Since this interaction is very complex, different optical units will benefit differently.

A. Without language model

The first set of experiments compares the performance of different optical units using the outputs of the optical model only. The recognition hypothesis is constructed by concatenating the best optical unit (corresponding to the output with the highest posterior probability) at each time-stamp. Table VI reports the performance for all tasks and on both development (Dev) and test (Test) datasets.

Results of the third row (i.e; Khatt-Random) are obtained with an optical model trained with the random paragraphs only. For the last three rows, the optical model is trained on both fixed and random paragraphs. In this case, results on both types of paragraphs are reported separately. It can be observed that:

First, for the Maurdor tasks (i.e; M-PRN and M-HWR), systems using optical units derived from isolated characters (Char and E−Char), generally, outperform those using optical units derived from character shapes (Shape and Sub−Char), although difference in the performance is not significant. This observation is also valid for the Khatt task using optical model trained and evaluated on the random paragraphs only (i.e; Khatt-Random in the table).

Second, results using optical model trained on the full Khatt database (fixed and random paragraphs) reveal an interesting behavior. When these models are evaluated on the fixed paragraphs only (i.e; Khatt-Fixed*), those using optical units derived from character shapes (i.e; Shape) perform the best. This is not the case, when they are evaluated on the random paragraphs only (i.e; Khatt-Random*). A possible explanation is that, the data contains a large set of examples (i.e; images) with the same text, in both training and evaluation dataset. This mismatch in the data, means that sufficient training examples were available for each shape of each character. In such training conditions, modeling the shape of the characters might lead to better results. This explains also the good performance on the fixed paragraphs.

B. With language model

The second set of experiments evaluate different optical unit sets when a language model is used. The goal is to search for the set of units that best interact with the lexical information provided by the language model. It is worth remembering that for a given task, the same hybrid 3−gram LM is used, independently on the optical units. Table VII reports systems performances on the Dev and Test datasets. From this table, it can be observed that:

First, adding language model component, performances are significantly improved. This is true for all optical unit sets. However, improvements on the Maurdor printed task (between 22.5% and 26.5%, relative) are less significant compared to the other tasks (between 38.2% and 41%, relative on the Maurdor handwritten task). This is can be expected, since recognizing printed text is less difficult than recognizing handwritten text. So, an optical model alone can perform well.

Second, the improvements with the (Char) optical units are more significant compared to those obtained with the other units. On the Maurdor handwritten, an improvement of 41% is obtained on the Dev dataset with the Char units. With the Shape and Sub−Char units, the improvements are about 38.2%. More interestingly, results on the Dev dataset of the fixed paragraphs of the Khatt task, shows that better improvement is obtained with the Char units (about 26.7%) compared to Shape units (about 24.2%), whereas the later units performed better without language model. As a result, the gape in the performance between Char and Shape optical units gets reduced.

Similar observations can be made for the Test datasets.

V. CONCLUSION

Optical models are the main component in state-of-the-art text recognition systems. Their role is to better model the relationship between the fundamental optical units and the text in the images. This work compared four Arabic optical unit sets. Such studies have already been conducted within the HMM framework. Given the success of the MDLSTM-RNN models, it was necessary to conduct such a study. The main findings of this work is that; First, optical models using isolated characters generally perform better, and they have the advantage of being the set with the smallest number of units. Second, when large amounts of training data is available, with a possible mismatch (in the lexical content) between train and evaluation data sets, it is preferable to use character shapes as basic units. Third, adding language model was more beneficial to the character units, and resulted in the best improvement.

For a fair comparison, optical models are trained from scratch and with the task-specific data only. As an extension to this work, we will investigate the effect of data augmentation techniques on different optical unit sets. This data can be used to train a large MDLSTM model, which will be used as a priori information for adaptation with task-specific data. Data augmentation can be also performed by using synthetic data or by transforming the task-specific data using some
image processing techniques [14]. Such transformations might improve (at the low cost) system performances.

**REFERENCES**


