

Techniques for Language Identification for Hybrid Arabic-English Document Images

Ahmed M. Elgammal
Department of Computer Science
University of Maryland
College Park, MD 20742, USA

Mohamed A. Ismail
Department of Computer Science
Faculty of Engineering
University of Alexandria, Egypt

Abstract

Because of the different characteristics of Arabic language and Romance and Anglo Saxon languages, recognition of documents written in hybrid of these languages requires that the language of the text to be identified priori to the recognition phase. In this paper, three efficient techniques that can be used to discriminate between text written in Arabic script and text written in English script are presented and evaluated. These techniques addresses the language identification problem on the word level and on textline level. The characteristics of horizontal projection profiles as well as runlength histograms for text written in both languages are the basic features underlying these techniques. Solving this problem is very important in building bilingual document image analysis systems which are capable of processing documents containing hybrid Arabic/Romance and Anglo Saxon languages.

1. Introduction

The development of multi-lingual document image analysis systems has become an important task with many applications recently. There is a big need for systems that are capable of handling documents with different languages. In the early 90s, several document analysis systems have appeared that are able to handle single language documents. There is a big need these days to expand document analysis systems to handle multi-lingual documents. In order to develop such systems we have to solve the problem of language identification.

This paper addresses the problem of language identification for documents printed in hybrid Arabic/English languages. Because of the many differences between Arabic and English text styles, character recognition methods differ in the way they handle text in these two languages. For example, Arabic is written from right to left while English is written from left to right. Also, Arabic text is cursive, i.e., characters are connected within each word and so Arabic OCR systems have an explicit or implicit segmentation phase to segment words into characters[1, 2, 3]. In contrast English characters are isolated and need no segmen-

tation unless if we consider the problem of touching characters due to low print quality [13, 6, 8]. Because of these basic differences, different approaches have been used by researchers for the OCR process for both languages. This makes it essential to identify the language priori to the OCR phase. Even if the same technique is used for solving the OCR problem, as in [3], still language identification priori to the OCR phase has the advantage of limiting the search space for the character recognition problem.

Multi lingual document analysis has many applications. one of the important applications is postal automation where envelopes have lines of different languages specially in the case of international mail. Also in digital libraries applications, technical journals and newspapers always have multi-lingual environment specially if the basic language of the document is not English. For example, in almost all technical journals and magazines that are written in Arabic, there are many words that are written in English as in figure 1. multi-lingual environment may also found in other documents such as maps and forms.

Language Identification prior to OCR has received some attention recently. In [10] multi-channel Gabor filtering is used to extract rotation invariant texture features that are used to identify the language of text blocks. In [9] features related to upward concavities in character structure are used to discriminate between two broad classes: western scripts and oriental (Korean, Japanese, Chinese) scripts. Different features were used to discriminate between languages in each of these classes. In [5] a combined analysis of several discriminating statistical features is used to discriminate between European and oriental language scripts.

In this paper, three techniques for language identification prior to OCR phase for hybrid Arabic-English documents are presented. These techniques may be generalized to include all other Romance and Anglo Saxon languages instead of only English and other languages that uses Arabic scripts such as Persian and Urdu. Although we emphasis on technical documents and magazines in the experiments, these techniques can be used in other document analysis ap-

plications.

The organization of the paper is as follow: Section 2 introduces the document analysis environment that we used. Section 3 presents three techniques for solving the identification problem. In section 4, experimental evaluation results of the proposed techniques are shown.

وتوجد آثار باقية كثيرة متناثرة في كل أرجاء المتزه ، بما في ذلك بقايا تماثيل شخصية بالية للملك مونتسوزوما الأول ، ومستشاره تلاكيليل . وبالقرب من البحيرة الأصفر يوجد مسرح Juventos Rosas المدرج ، وتقام فيه الحفلات الموسيقية العامة . وهو نقطة بداية شارع الشعراء حيث تماثيل Antonio Plaza - Sor Juana Ines de la cruz ، Manuel Acuna ، - التصفية العملاقة ، إلى جوانب تماثيل آخرين ، وتقوم على قواعد حجرية . كما يشاهد المثلث النسخ من الفترة قبل الإسبانية عند مدخل Anthropology المتسحف القومي لعلم الإنسان) . وعمود الطورم الخشبي المنقوش الذي تبرعت به الحكومة الكندية ، والمعبد متعدد الأروار الذي أعدته جمهورية كوريا للسكيبك في سنة ١٩٦٨ .

هذه الفصيلة عدداً من الطيور تختلف عن بعضها ظاهرياً فقط: فهناك ، منجليات المنقار ، sicklebills ، و ، الهزوتيات ، parotias و ، الأستراليات طويلة الذيل ، long-tailed astrapias و ، طيور الجنة كثيفة الريش ، heavily plumed ، paradisaeas و ، المانوكيدات الزرقاء السوداء ، blue-black manucodes . وتتصف كل هذه الطيور بصلابة البنية وقوة الأرجل والأقدام ، إلا أنها تختلف من نوع إلى نوع من حيث اللون والريش . إن معظم الأنواع ذات هيئة جنسية ثنائية ، sexually dimorphic ، أي يختلف مظهر الذكر اختلافاً واضحاً عن الإناث ، فعالمنا ما تكون الذكور زاهية الألوان ، ذات ريش طويل متركب يسمى الريش العُزسي ، nuptial plumes ، في حين تغتفر الإناث إلى الريش المتخصص ، ويغلب عليها دائماً اللون البني أو الأسود .

Figure 1. Examples of hybrid Arabic-English language environment

2. Document Analysis Environment

The problem that we are concerned about in this paper can be described as follows: Given an image of a word or a collection of words in one text line, how can we determine what language it is written in prior to solving the OCR problem provided that it is written in one language. We consider the case of documents written in hybrid Arabic/English environment as those in figure 1. The proposed techniques are part of our research bilingual (Arabic/English) document analysis system.

The approach that is used for document analysis affects the technique used for language identification. There are two basic approaches for document analysis [4]; top down approach and bottom up approach. In the top down approach, starting with the scanned page, the page is segmented into large blocks which are then classified to text blocks, Figures, tables, etc. Text blocks are then resegmented into text lines and then into words and characters. In the bottom up approach, black pixels are grouped into small components (character size). These components are combined to form words and then text lines and paragraphs. We use a bottom up approach to analyze the document similar to that used by Tsujimoto and Asada [11]. First, connected component extraction is performed on the thresholded image after horizontal smearing [4]. Connected components are grouped together horizontally through a series of steps to form a tree structure of subwords, words, lines and paragraphs. Language identification is performed on word level,

i.e., after extracting tokens corresponding to words, these tokens are passed to a language identification process to label them either Arabic or English.

The techniques proposed in section 3 are general enough to solve the problem of Arabic/English language identification on two levels:

- Textline level: In this case, text lines are either written in Arabic or in English. Consider, for example, postal automation applications that handle international mail where you can find complete text lines written in Arabic or in English.
- Word level: We can call this case, hybrid Arabic-English writing. where the same text line contains both Arabic and English words. This is the basic problem that we are going to emphasize on.

Other document analysis applications require solving the language identification problem on the page level or on the paragraph level where pages or paragraphs are written in the same language.

3. Techniques for Language Identification

3.1. Horizontal Projection Profile

The black-count horizontal projection profile is a one-dimensional integer-valued function $f(y)$, where the value of f is the number of black pixels in row y . Projection profiles have been used extensively in the field of document analysis specially in skew removal [4] and for block classification [12]. Fig. 2 illustrates some Arabic and English text lines and their horizontal projection profiles.

As can be noticed from these figures, there are different characteristics for the horizontal projection of each of the two languages. Two basic characteristics are useful in discriminating between the two languages:

- The horizontal projection profiles of Arabic text have a single peak around the middle of the text line. This peak corresponds to the baseline of the Arabic writing where characters are connected together. In contrast, projections of English text have two major peaks.
- The projections of Arabic text lines are smooth while the projections of English text line have sharp jumps.

Two techniques are described and evaluated below to discriminate between these two types of projection profiles.

3.1.1 Peak Detection

In this method the peaks in the horizontal projection profile are detected. First the projection is normalized with respect to the total length of the profile. Fig. 3 illustrates a typical normalized profile for a typical Arabic and English text line.

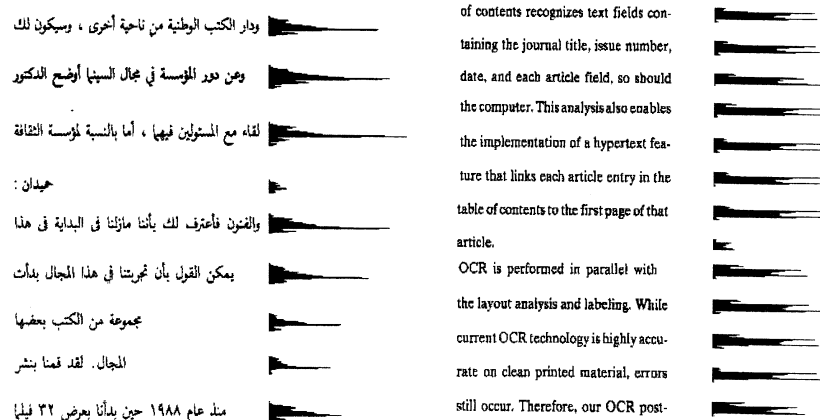


Figure 2. On the left, sample Arabic text lines and their horizontal projection profiles. On the right, sample English text lines and their horizontal projection profiles.

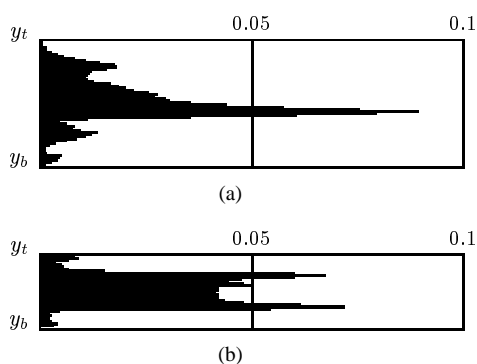


Figure 3. The normalized projection profiles (a) a typical normalized profile for Arabic text line. (b) a typical normalized profile for English text line. The numbers present the ratio of the height to the total length. The total length is 1. y_t and y_b are the top and bottom of the text line

The profiles are smoothed by convolution with a Gaussian kernel. Peaks that have heights more than certain threshold are counted and their location are considered. Here we explicitly utilize the first observation about horizontal projection profiles characteristics as was mentioned above, that is, projections of Arabic text lines have a single peak around the middle of the line while projections of English text lines have two peaks, one in the upper half of the line and one in the lower half.

According to the experimental results that will be shown in section 4, a simple technique as peak detection gives excellent results in case of text line-level language identification. In contrast it gives lower performance in case of word-level language identification. That is because in the

latter case, a word's projection does not preserve the characteristic of peaks as complete text lines do. A single word's profile is always sensitive to the characters in the word specially if the word is too short in length and have small number of characters.

3.1.2 Use of Moments

As mentioned before, the horizontal projection profile represents the accumulated runlength in the horizontal direction. This technique is based on finding the moments of the horizontal projection profile and use them as features in discriminating between Arabic and English text. The moments in the dimensionless form are used to avoid variations in the heights of the profile due to variations in text line length.

Let $f(y)$ to be the height of the horizontal projection profile at row y where $y_{top} \leq y \leq y_{bottom}$. then the r th moment about the mean is

$$m_r = \frac{\sum_{y=y_{bottom}}^{y_{top}} (f(y) - \bar{h})^r}{y_{bottom} - y_{top} + 1}$$

where

$$\bar{h} = \frac{\sum_{y=y_{bottom}}^{y_{top}} f(y)}{y_{bottom} - y_{top} + 1}$$

the r th moment in the dimensionless form is

$$a_r = \frac{m_r}{s^r} = \frac{m_r}{(\sqrt{m_2})^r}$$

where $s = \sqrt{m_2}$ is the standard deviation. Since $m_1 = 0$ and $m_2 = s^2$, we have $a_1 = 0$, $a_2 = 1$.

Fig. 4 illustrates the values of the third, fourth and fifth moments for a group of Arabic and English text lines with different length. The moments in the case of English text lines tends to be smaller than those of Arabic text lines. A

| | | | |
|---|-------|-------|--------|
| ودار الكتب الوطنية من ناحية أخرى ، وسيكون لك | 1.656 | 5.306 | 14.530 |
| وهن دور المؤسسة في مجال السينما أوضح الدكتور | 1.428 | 4.396 | 10.803 |
| لقاء مع المسؤولين فيها ، أما بالنسبة لمؤسسة الثقافة | 2.072 | 6.947 | 22.029 |
| حميدان : | 0.570 | 2.846 | 3.611 |
| والفنون فأعترف لك بأننا مازلتنا في البداية في هذا | 1.622 | 5.398 | 15.070 |
| يمكن القول بأن نمرتنا في هذا المجال بدأت | 1.692 | 5.314 | 14.471 |
| مجموعة من الكتب بعضها | 1.610 | 4.720 | 12.478 |
| المجال. لقد قمنا بنشر | 2.104 | 7.017 | 21.825 |
| منذ عام ١٩٨٨ حين بدأنا بعرض ٢٢ فيلماً | 1.010 | 3.209 | 6.532 |
| تتلو | 1.932 | 6.120 | 17.529 |
| بتصل بتطور التعليم عننا ، وبعضها « دراسات في | 1.983 | 6.492 | 19.695 |
| سينمائيًا في هذا العام بعد أن نجحنا في صياغة علاقات | 1.596 | 4.840 | 12.927 |
| الثقافية وثقافة الإمارات ، وبعضها « عن الشعر | 1.421 | 4.306 | 10.514 |
| فنية إيجابية مع الجهات المعنية بالفن السينمائي في | 1.643 | 5.345 | 15.084 |
| المنطقة الخليج والجزيرة ، وعن تحليل القيم | 1.615 | 5.012 | 13.480 |

| | | | |
|--|--------|-------|--------|
| than the scan region, the background | 0.271 | 1.474 | 1.156 |
| the amount of data (thus, the process- | 0.259 | 1.454 | 1.104 |
| outside the journal often contains this | 0.198 | 1.504 | 1.026 |
| Image and document | 0.275 | 1.417 | 1.145 |
| ing time) and the artifacts that might | 0.278 | 1.680 | 1.544 |
| noise. The gutter region along the spine | 0.238 | 1.472 | 1.046 |
| otherwise impede the recognition of true | 0.213 | 1.387 | 0.868 |
| of thickly bound journals can also yield | 0.128 | 1.412 | 0.663 |
| processing | -0.270 | 1.535 | -0.443 |
| features. | -0.191 | 1.942 | -0.181 |
| noise. Whatever its cause, we wish to | -0.135 | 1.334 | -0.204 |
| Document layout analysis and logical | 0.188 | 1.505 | 1.011 |
| reduce noise from this initial point of | 0.133 | 1.513 | 0.914 |
| labeling are then performed on the | 0.258 | 1.582 | 1.212 |
| processing as much as possible to mini- | 0.200 | 1.445 | 1.053 |

Figure 4. Values of the third, fourth and fifth moments for some Arabic and English text lines

two layer feed forward neural network is used as a classifier in this problem. The input to the network is the three moments m_3, m_4, m_5 , i.e., each input node is a continuous valued input. Four nodes are used in the hidden layer and two nodes for the output (Arabic/English). The back propagation [7] algorithm is used to train the network. The back propagation training algorithm is an iterative gradient algorithm designed to minimize the mean square error between the actual output and the desired output.

According to the experimental results that will be shown in section 4, The moment method gives excellent performance in both the case of textline-level language identification and the case of word-level language identification.

3.2. Runlength Histogram

In this method each runlength, (x_s, x_e, y) , where x_s and x_e are the start and the end points of the runlength and y is its horizontal location, is mapped to a pair (loc, len) where loc is the normalized vertical location with respect to the height of the textline and len is the normalized length, i.e.,

$$loc = \frac{y - y_t}{y_b - y_t}$$

$$len = \frac{x_e - x_s}{y_b - y_t}$$

The distribution of runlengths over the 2-D location-length space is a very discriminating feature to discriminate Arabic and English text. Figure 5 shows an example for the 2-D location-length plot for runlengths of an Arabic and an English text line. The 2-D histogram of the runlengths in a given text line is constructed by dividing the (loc, len) space horizontally and vertically into bins and the count of

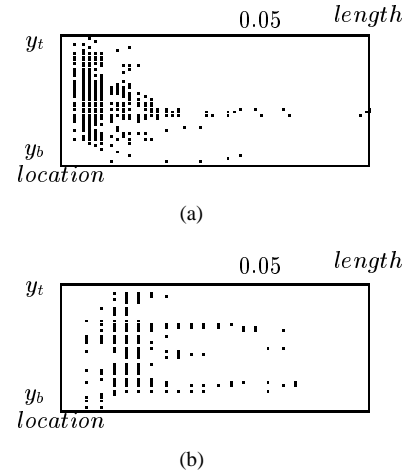


Figure 5. Runlength location-length 2-D plots. (a) for a typical Arabic text line. (b) for a typical English text line.

the runlengths in each bin is calculated. We used 8x8 bins for the histogram and the histogram is then normalized. A two layer feed forward neural network is used as a classifier. The input layer contains 64 nodes taking as an input the histogram normalized bin count. We used 10 hidden nodes and 2 output nodes. The sigmoid function is used as output activation function for the output and hidden nodes. The network is trained using the backpropagation algorithm with text lines and words of various length.

4. Experimental Results

The suggested methods for language identification are tested and evaluated using sample documents. The testing

were performed on both textline-level and word-level in order to enable us to evaluate the performance of each technique.

In the Experiments, three sets of documents were used. The first set contains Arabic documents from 4 different Arabic magazines and also contains pages written using word processors and printed on laser printer. This variety of sources in the Arabic set is to ensure the existence of variety of fonts and sizes in the sample. The second set of documents contains English documents from 2 different magazines (*IEEE computer* and *IEEE JSAC*.) The third set contains documents with hybrid environment (mixed Arabic and English text) some of them from Arabic magazines that contains English text and some of them were prepared using a word processor and printed on laser printer. All the documents were scanned with resolution 300 dpi and processed to separate text lines and words that are units of evaluations. The first and second sets were used to obtain training samples for Arabic and English text respectively, The training set contains about 12,000 text lines of various lengths ranging from very short words to full text lines.

For textline-level evaluation, 816 Arabic textlines and 1160 English textlines of various lengths were used as a test set. Table 1 shows the results obtained for each of the techniques described in section 3 evaluated on text line level. For word-level evaluation, 8320 words (4168 Arabic and 4152 English) were used as a test set. Table 2 shows the results obtained for each of the techniques described in section 3 evaluated on word level. As can be noticed from the results in both cases, the three techniques give good results in the case of textline level language identification. However, the runlength histogram is the most robust in the case of word level language identification.

| Textline-Level Language Identification | | | | | |
|--|----------|---------|------|------|-------|
| Method | Language | Samples | Hit | Miss | Ratio |
| Peak Detection | Arabic | 816 | 775 | 41 | 95% |
| | English | 1160 | 1089 | 71 | 93.9% |
| Moments | Arabic | 816 | 800 | 16 | 98% |
| | English | 1160 | 1156 | 4 | 99.7% |
| Runlength Histogram | Arabic | 816 | 798 | 18 | 97.8% |
| | English | 1160 | 1143 | 17 | 98.5% |

Table 1. Results for textline-level language identification

| Word-Level Language Identification | | | | | |
|------------------------------------|----------|---------|------|------|-------|
| Method | Language | Samples | Hit | Miss | Ratio |
| Peak Detection | Arabic | 4168 | 2876 | 1292 | 69% |
| | English | 4152 | 3155 | 977 | 76% |
| Moments | Arabic | 4168 | 3856 | 312 | 92.5% |
| | English | 4152 | 3852 | 300 | 92.8% |
| Runlength Histogram | Arabic | 4168 | 4006 | 162 | 96.1% |
| | English | 4152 | 4019 | 133 | 96.8% |

Table 2. Results for word-level language identification

5. Conclusion

We presented three simple and efficient techniques to discriminate between words and text lines written in Arabic and English. The three techniques utilize the different characteristics of Arabic and English text. The characteristics of the horizontal projection profiles as well as runlength histograms are used as features for the classification. We presented an approach based on detecting the peaks in the horizontal projection profile. We presented another approach based on the moments of the profiles using neural networks for classification. Finally, we presented an approach based on classifying runlength histogram using neural networks. We achieved a correct classification of 99.7% for text line level language identification and 96.8% for word level language identification.

References

- [1] A. Amin. Arabic character recognition. In H. Bunke and P. Wang, editors, *Handbook of Character Recognition and Document Image Analysis*, pages 397–420. World Scientific Publishing Company, 1997.
- [2] A. Amin. Off line arabic character recognition- a survey. In *ICDAR*, pages 596–599, 1997.
- [3] I. Bazzi, R. Schwartz, and J. Makhoul. An omnifont open-vocabulary ocr system for english and arabic. *IEEE PAMI*, 21(6):495–504, June 1999.
- [4] R. G. Casey and K. Y. Wong. Document-analysis systems and techniques. In R. Kasturi and M. M. Trivedi, editors, *Image Analysis Applications*, pages 1–36. Marcel Dekker, Newyork, 1990.
- [5] J. Ding, L. Lam, and C. Y. Suen. Classification of oriental and european scripts by using characteristic features. In *ICDAR*, 1997.
- [6] J.Mantas. An overview of character recognition methodologies. *Pattern Recognition*, 19:425–430, 1986.
- [7] R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP*, pages 4–22, Apr. 1987.
- [8] S. Mori, C. Y. Suen, and K. Yamamoto. Historical review of ocr research and development. *Proceedings of the IEEE*, 80(7):1029–1058, July 1992.
- [9] A. L. Spitz. Determination of the script and language content of document images. *IEEE PAMI*, 19(3), Mar. 1997.
- [10] T. Tan. Rotation invariant texture features and their use in automatic script identification. *IEEE PAMI*, 20(7), July 1998.
- [11] S. Tsujimoto and H. Asada. Major components of a complete text reading system. *IEEE proceedings*, 80(7):1133–918, July 1992.
- [12] M. Viswanathan and G. Nagy. Characteristics of digitized images of technical articals. Technical report, IBM Storage Products Division and Rensselaer Polytechnic Institute, 1992.
- [13] V.K.Govindan and A. Shivaprasad. Character recognition — a review. *Pattern Recognition*, 23:671–683, 1990.