Literary and Linguistic Computing Advance Access published October 1, 2014

The influence of language orthographic characteristics on digital word recognition

Ofer Biller, Jihad El-Sana and Klara Kedem
Ben-Gurion University of the Negev, Israel

Correspondence:
Ofer Biller, Department of Computer Science, Ben Gurion University of the Negev, P.O.B 653 Be'er Sheva 84105, Israel
E-mail:
billero@cs.bgu.ac.il

Abstract

This research studies the effect of language orthographic characteristics on the performance of digital word recognition in degraded documents such as historical documents. We provide a rigorous scheme for quantifying the statistical influence of the orthographic characteristics on the quality of word recognition in such documents. We study and compare several orthographic characteristics for four natural languages and measure the effect of each individual characteristic on the digital word recognition process. To this end, we create synthetic languages, for which all characteristics, except the one we examine, are identical, and measure the performance of two word recognition algorithms on synthetic documents of these languages. We examine and summarize the influence of the values of each characteristic on the performance of these word recognition methods.

1 Introduction

Research in digital script recognition could be classified to two main approaches: segmentation-based and segmentation-free recognition. The segmentation-based approach segments an input word into individual characters, which are then recognized and combined to identify the input word. The segmentation-free approach (the holistic approach) recognizes a whole word without segmenting it (opposed to classical Optical Character Recognition (OCR)). Recently, the holistic approach for digital word recognition has been attracting more interest and has become widely accepted in the handwriting recognition research community, e.g. Plamondon and Guerfali (1996), Plamondon and Srihari (2000), Gatos *et al.* (2005), Biadsy *et al.* (2006), Zagoris *et al.* (2006). This approach has been adapted for word recognition of historical documents which usually suffer from high level of degradation (Lavrenko *et al.*, 2004; Rath and Manmatha, 2007).

In the holistic approach, the quality of the results depends on characteristics of the language. For example, a word for which there are many words with similar characters and length in the lexicon, may have a higher chance of being misclassified. Such an observation calls for examining the relation between language characteristics and the quality of automatic word recognition. The influence of several language orthographic characteristics on visual word recognition has been studied in the domain of cognitive psychology perception (Richards and Heller, 1976; Coltheart *et al.*, 1977; Andrews, 1989, 1992, 1997; Grainger, 1990; Perea and Rosa, 2000; New *et al.*, 2006). In this research, we explore the influence of these characteristics on automatic word recognition for low quality and degraded documents, such as historical documents.

We examine the following orthographic characteristics: word length, word orthographic neighborhood, and distribution of ascenders and descenders among text characters. Intuitively, these

doi:10.1093/llc/fqu051

characteristics statistically influence the quality of holistic recognition methods. Here we provide a rigorous scheme for measuring and validating this intuition.

We start with learning the behavior of these characteristics in natural languages, by gathering text documents in English, Hebrew, Arabic, and Russian and analyzing the distribution of these characteristics in them.

To isolate the effect of a specific characteristic, we generate synthetic languages where all characteristics are identical except the one we examine. Then we produce document images for these languages, apply word recognition algorithms, and analyze the relations between each language characteristic and the performance of the recognition. We simulate degraded documents by adding random noise to the synthetically generated documents in a controlled manner.

The analysis of the relationship between language characteristics and the performance of recognition algorithms can determine the suitable recognition approach for a given language, and simplify the utilization of existing techniques. This is useful in the research of degraded documents in different languages. Here we examine two approaches for word spotting (recognition): dynamic time warping (DTW) offered by Rath and Manmatha (2003b) and gradient-based binary features (GSC) used in Zhang *et al.* (2004).

The article is organized as follows. In Section 2, we outline related work. In Section 3, we present the language characteristics and review their effect on human word recognition as reported in works in psychology and cognitive research. In Section 4, we show the distribution of these characteristics for several natural languages. Section 5 measures the influence of each characteristic on word recognition, using the synthetically generated languages and degraded documents. Finally we draw conclusions in Section 6.

2 Related Work

The relation between language characteristics and word recognition has attracted the interest of researchers in the field of psychology and cognitive research. The orthographic similarity (orthographic neighborhood) was shown to affect human performance of word recognition, e.g. Grainger (1990), Andrews (1992, 1997), Perea and Rosa (2000). The commonly used definition for word neighborhood was introduced by Coltheart *et al.* (1977) as the set of words that can be received by replacing a single letter in the original word. Recently, a more flexible definition was suggested, which is based on Levenshtein's string distance metric (Levenshtein, 1966; Yarkoni *et al.*, 2008). The new method takes into account letter replacement, but also insertion and deletion. Word length has also been shown to have influence on visual word recognition (Richards and Heller, 1976; New *et al.*, 2006).

The segmentation-free holistic approach to script recognition has become more popular in recent years, e.g. Madhvanath and Govindaraju (2001). This approach has been adapted for word recognition of historical documents, which usually suffer from high level of degradation (Lavrenko *et al.*, 2004). Spitz (1999) makes use of the presence of ascending and descending letters for both recognition and for language identification (Spitz, 1997). Some work has been done on image quality measurement and OCR result prediction, e.g. Esakov *et al.* (1994); Blando *et al.* (1995); Ye *et al.* (2012); Salah *et al.* (2013), but these mostly focus on OCR while our work deals with holistic word recognition.

3 Natural Language Characteristics

Below we present the language characteristics we use and our hypotheses on their influence on the performance of word recognition. The characteristics we examine are *word length*, *orthographic neighborhood*, *distribution of ascenders and descenders*. In a preliminary stage we examined a number of possible characteristics and decided to focus on the characteristics above due to their high influence on the structural shape of words. Throughout the research article, we use the same characteristics for all the examined languages (as described below).

3.1 Word length

We expect word recognition algorithms to perform better in languages with longer words. We consider the averages of word length in the languages as an orthographic characteristic.

3.2 Orthographic neighborhood

This characteristic aims to measure orthographic distribution of words in the language. The assumption is that low orthographic distance between words is manifested in small distance between word images which contributes to lower recognition rates. An accepted measure for this is the *orthographic neighborhood*. In this article, we use Levenshtein distance for determining the orthographic neighborhood. We take the average neighborhood size over all language words as a representative measure.

3.3 Ascenders and descenders

An important aspect that has high impact on the quality of word recognition is the graphical properties of the character set. We focus on ascenders and descenders. Graphically a handwritten line has a lower baseline and an upper baseline, between which most letters are written. Some letters exceed these baselines. Those that exceed the upper baseline are called ascenders (e.g. b, d). The ones that exceed the lower baseline are called descenders (e.g. g, q). Ascenders and descenders influence dramatically the outer shape of the word's image, therefore, might effect word recognition. Several approaches use the presence of ascenders and descenders for recognition and language identification, but to the best of our knowledge, no work has been done to quantify this influence.

4 Distribution of Orthographic Characteristics in Natural Languages

Here we analyze the behavior of the characteristics described above in several natural languages. We collected a corpus of books in four different languages: English, Russian, Hebrew, and Arabic. The books we used were a random collection of prose books in text format (some of which were—Adventures of Huckleberry Finn, Crime and Punishment, and variety of short stories which were available in several languages). We processed the text by extracting distinct words and building a dictionary for each language. We prefer collecting word statistics from documents and not from dictionaries in order to get a better representation of frequently used words and to ignore rare and unused words. For each of the languages, our database contains about 18,000 unique words.

4.1 Word average length

Figure 1 shows that both English and Russian have on average longer words than Hebrew and Arabic and a wider spread of word lengths. The graph in Figure 1 illustrates the distribution of word length by showing the percentage of words, for each length over all words in the corpus. It shows substantial differences among the examined languages.

4.2 Word orthographic neighborhood

To evaluate the orthographic density of the words in each language, we calculate for each word in a language the number of its neighbors (within the corpus). The graph in Figure 2 shows for each language the percentage of words having a specific neighbor count (the x-axis). As can be seen, in English and Russian the main mass of words concentrates in the lower values of neighbors per word, while Hebrew and Arabic spread also over higher values of neighbors.

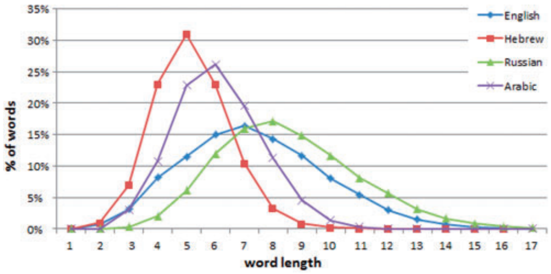


Fig. 1 Distribution of word length per language—percentage of words per word length

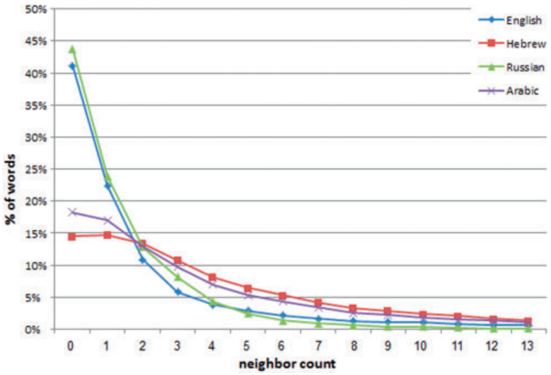


Fig. 2 The distribution of word neighbors count—shows the percentage of words having specified neighbors count, for English, Hebrew, Russian, and Arabic

Table 1 Ascender and descender configurations—for each language, the number of non-empty configuration sets and the average number of words in a configuration set

Language	Non-empty configurations	Average configuration size
English	4,218	4.29
Russian	2,742	6.87
Hebrew	499	39.59
Arabic	1,560	11.77

4.3 Ascending and descending letters

To measure the distribution of ascending and descending letters, we define *ascender and descender configuration*. Each configuration is represented by a sequence of letter types (ascending/descending/none). Each word matches some configuration. For example the word 'dog' will match the configuration ascend-none-descend. For each language, we compute how many words are in a given configuration set. In Table 1 we present the average size of the configuration sets in each language. A configuration is considered empty for a language if the language does not contain any words of that configuration. We expect languages with many small non-empty configuration sets to have more graphical diversity in words' general shapes and therefore to have an advantage in word recognition.

4.4 Summary of the results for the characteristics on the four natural languages

As seen above there are substantial differences in the values of the checked characteristics amongst the languages. In general, English and Russian have longer words and lower number of neighbors per word than Hebrew and Arabic. As to ascender and descender distributions, Hebrew words have the lowest diversity, and English has the widest spread among the examined languages.

5 Experimental Evaluation

In previous sections, we listed several natural language characteristics which we conjecture influence the quality of word recognition. We measured some of these characteristics on corpora of four natural languages and found out that there are substantial differences between these characteristics among different languages. Yet, we have not given any estimation or measurement as to the influence of each of these characteristics on word recognition. Using real documents in different languages does not enable the isolation of the effect of each characteristic. To measure the impact of a certain characteristic of natural languages, we generated a set of synthetic languages, based on the English alphabet and similar in every characteristic but the examined one. For the examined characteristic, the languages have different values. For each language we generate documents, perform word spotting, evaluate the results and compare their accuracy. Because the generated languages and documents are similar in every aspect except the examined characteristic, a distinctive and stable variance in test results would be confidently assigned to the examined characteristic.

On clean, high quality, synthetic documents we would receive near perfect word recognition results regardless of the values of the language characteristics. Therefore, we added noise to the documents to create variance in the quality of test results. We used few different combinations of degradation types and levels such as simple random Gaussian and salt & pepper noise, and Kanungo's degradation model (Kanungo *et al.*, 1993). In all experiments,

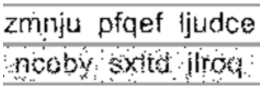


Fig. 3 Two examples of a noised document segments from the conducted experiments. The examples use different degradation models, the upper segment is degraded using Gaussian and salt & pepper noise, and the lower is degraded using Kanungo's model. The same degradation and binarization processes were applied on all generated documents

the reaction to change in the characteristics was similar across different types of degradations. The results provided are the averages over the used degradation models. Example of the generated noise is shown in Figure 3. The degradation models' parametrization was adjusted to receive mediocre recognition results, and stayed constant throughout all experiments. Among the many available approaches for word recognition, we chose to start with DTW based approach, presented by Rath and Manmatha (2003b) with the combination of profile features from Rath and Manmatha (2003a), and the GSC method (Zhang *et al.*, 2004).

5.1 Evaluation of the word spotting

As part of document generation, we construct a ground truth data base on which we test our recognition results. Using both word recognition methods, we retrieve all the occurrences of each word in the corpus from the generated documents. The overall performance for a given method on a given language is the average F-Measure for all the retrieved words.

5.2 The testing system

We have built a testing system that runs sets of tests. Each set is composed of tests with similar languages which differ only in the examined characteristic. The process for one test includes creation of a language and a noised document in this language, running two word recognition methods, and evaluating the recognition results for both. The whole process of running a test set is automatic, and can be run several times in order to increase the reliability of the results. Test specifications and test results are written to a data base, enabling fast creation of new tests and easy access to test results. The testing software is very robust and enables conducting tests on other characteristics in the future.

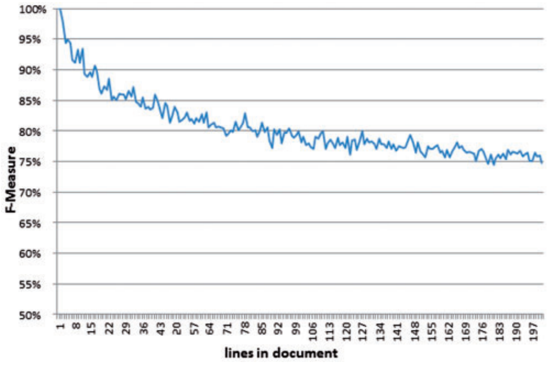


Fig. 4 Stability calibration graph—test grade (F-Measure) as function of line count in document

5.3 Stability calibration

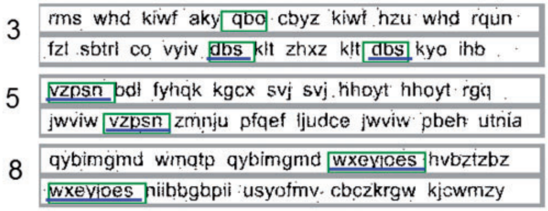
The process of data generation includes many random elements such as word generation in the synthetic language dictionary, word selection for the synthetic document, and degrading the document. A set of predefined rules and parameters control these random elements, yet they have some influence on the results. The larger the volume of text in the test, the smaller the effect of the random elements is on the results. Therefore, we had to find the sufficient size of tested documents in order to reduce the random factor in the results. We ran the system on inputs consisting of a growing number of lines and plotted the corresponding F-Measure. As seen in Figure 4, as the number of lines in the document increases, the F-Measure stabilizes.

To evaluate the fluctuation of the results, we calculate the standard deviation of the F-Measure for each set of five subsequent tests. As the number of text lines used increases, the standard deviation of the subsequent test results decreases. This evaluation shows that above 150 lines of text, the standard deviation stays below 1%. According to this, we set the number of words per document to 2000 for all tests, i.e. over 200 text lines in a document.

5.4 Experiments and results

Next we describe the experiments performed on the test system for each of the examined language characteristics.

Average word length: We generated nine languages, each with different word length average, between 2 and 10. The standard deviation of word lengths was set to one for all languages. The length of each word in a test set was randomly generated with normal distribution of the specified average and standard deviation. The experiment was repeated three times for validating the results. Figure 5 shows two sample lines from three documents from three languages with word length averages of 3, 5, and 8, respectively. Figure 6 summarizes the results of the word length experiment. It displays the F-Measure of word recognition as a function of word length average, for DTW and GSC methods. We notice that the difference between the quality of the results of both methods is high. The F-Measure for a language with the short words was very low for DTW, and above 80% for GSC. As the average word length increases, the F-measure for DTW approaches 100%, while F-Measure decreases for GSC for average word length 8 and above. We believe that this is because GSC was fine-tuned for English, where average word length is 7, as seen in Figure 1



3
rms whd kiwf aky qbo cbyz kiwf hzu whd rqun
fzl sbtrl co vyiv dbs klt zhxz klt dbs kyo ihb

5
vzpsn bdl fyhqk kgcx svj svj hhoyt hhoyt rgq
jwviw vzpsn zmnju pfqef ljudce jwviw pbeh utnia

8
qybimgmd wmqtp qybimgmd wxeyioes hvbztzbz
wxeyioes niibbgbpii usyofmv cbczkrgw kjcwmzy

Fig. 5 Examples of spotting words in noisy documents (by DTW). Two example lines are from three documents with average word lengths of 3, 5, and 8. The query words are marked with an underline, and the matches found by the system are marked with a rectangle (therefore the word in the first line marked with a rectangle and not underlined is an example of false positive)

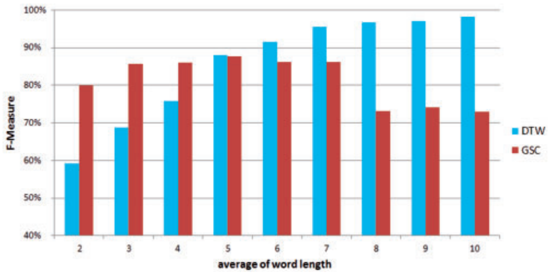


Fig. 6 Word length experiment—quality of the results (F-Measure) as a function of average word length in a language

Ascending and descending characters: This experiment examines the influence of ascender and descender distribution on word recognition results. For that purpose, we create randomly synthetic test languages using a partial set of the English lower case letters. The alphabet of each test language consists of 14 characters with a different composition of ascenders and descenders. We used more than one test language for each composition to eliminate the influence of the choice of a specific character set. Figure 7 shows the average F-Measure for each combination (#ascenders_#descenders_ #normal)

The DTW with the projection profile features shows significantly better results for balanced combination of ascenders and descenders while the GSC method is less influenced by this characteristic. For both recognition methods the results are similar with F-measure around 75 and 80%, while the recognition by DTW for words with no ascenders and descenders is much lower (about 65.5%).

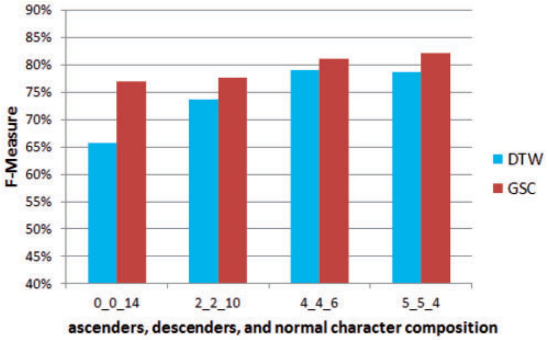


Fig. 7 The graph depicts the F-Measure as a function of ascender and descender composition in a language

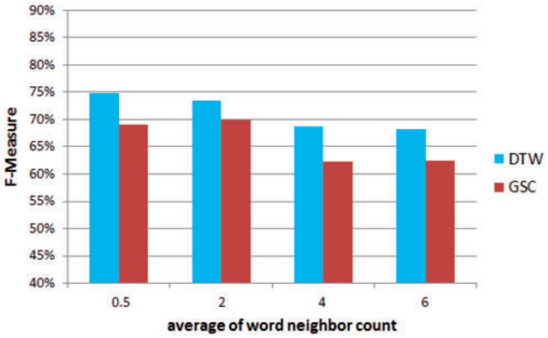


Fig. 8 F-Measure results as a function of average neighbor count

Orthographic neighborhood: In this experiment, we created documents with different average neighbor count for the words (and fixed average word length to be 7). The created test set contained the values of 0.5, 2, 4, and 6 as average word neighbor count. The graph in Figure 8 displays the F-Measure results for DTW and GSC, and shows that in both methods the quality of recognition declines as the number of the neighboring words grows.

6 Conclusions and Future Research

In this research, we examine the effect of several orthographic language characteristics on word recognition. We examine word length, orthographic neighborhood, and distribution of ascenders and descenders. These characteristics show very different behaviour in the natural languages we tested; e.g., Hebrew and Arabic have on average shorter words than English and Russian, a larger number of orthographic neighbors, and also fewer ascender–descender configurations.

We investigated the influence of these characteristics on the performance of word recognition in degraded documents, by running word spotting tests on synthetically generated languages (all the synthetic languages were constructed of the English alphabet). The synthetic languages are built so that just the examined characteristic varies while all the others remain the same.

We conclude that for the characteristics ascender/descender and for average neighbor count DTW and GSC demonstrate a similar behavior with minor differences, in which low number of neighbors per word and balanced distribution of ascenders and descenders contribute to better results of word recognition. On the other hand, for average word length, DTW clearly presents better results as words get longer, while GSC responds best when average word length is between 3–7 letters. This can be explained by the fact that DTW allows flexibility of word length (Rath and Manmatha, 2007), while GSC performs substitution of words to a predefined number of regions within the word frame (Zhang *et al.*, 2004), which leads to optimal behavior in a limited range of word lengths.

Acknowledgements

This research was supported in part by the DFG-Trilateral Grant no. 8716, the Lynn and William Frankel Center for Computer Sciences, and the Paul Ivanier Center for Robotics and Production Management at Ben-Gurion University, Israel.

References

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**: 802–14.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**: 234–54.

Andrews, S. (1997). The effects of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychological Bulletin and Review*, **4**: 439–61.

Biadsy, F., El-Sana, J., and Habash, N. (2006). *Online Arabic handwriting recognition using Hidden Markov Models*. Proceedings of the 10th International Workshop on Frontiers of Handwriting and Recognition, La Baule, Centre de Congres Atlantia, France. pp. 1009–33.

Blando, L. R., Kanai, J., and Nartker, T. A. (1995). *Prediction of OCR accuracy using simple image features*. Montreal, Canada: ICDAR, pp. 319–22.

Coltheart, M., Davelaar, E., Jonasson, J. F., and Besner, D. (1977). Access to the internal lexicon. In Dornic, S. (ed.), *Attention and Performance VI*. Hillsdale, NJ: Erlbaum, pp. 535–55.

Esakov, J., Lopresti, D. P., and Sandberg, J. S. (1994). *Classification and Distribution of Optical Character Recognition Errors, Proceedings of SPIE — The International Society for Optical Engineering*, Orlando, Florida, *2181*, pp. 204–16.

Gatos, B., Konidaris, T., Ntzios, K., Pratikakis, I., and Perantonis, S. (2005). *A Segmentation-Free Approach for Keyword Search in Historical Typewritten Documents, Proceedings of Eighth International Conference on Document Analysis and Recognition*, **Vol.1**. Seoul, Korea, pp. 54–58.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, **29**: 228–44.

Kanungo, T., Haralick, R. M., and Phillips, I. T. (1993). *Global and Local Document Degradation Models* ICDAR, Tsukuba City, Japan: IEEE, pp. 730–4.

Lavrenko, V., Rath, T. M., and Manmatha, R. (2004). *Holistic Word Recognition for Handwritten Historical Documents, Document Image Analysis for Libraries*. Palo. Alto, CA, pp. 278–87.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, **10**: 707.

Madhvanath, S. and Govindaraju, V. (2001). The role of holistic paradigms in handwritten word recognition. *IEEE Transactions Pattern Analysis Machine Intelligence*, **23**(2): 149–64.

New, B., Ferrand, L., Pallier, C., and Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the english lexicon project. *Psychonomic Bulletin and Review*, **13**(1): 45–5.

Perea, M. and Rosa, E. (2000). The effects of orthographic neighborhood in reading and laboratory word identification tasks: A review. *Psicológica*, **21**: 327–40.

Plamondon, R. and Guerfali, W. (1996). *Why Handwriting Segmentation can be Misleading? Proc. Intl Conf. on Pattern Recognition*. Vienna, Austria, pp. 369–400.

Plamondon, R. and Srihari, S. N. (2000). On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Transactions Pattern Analysis Machine Intelligence*, **22**: 63–84.

Rath, T. and Manmatha, R. (2003a). *Features for Word Spotting in Historical Manuscripts. Proceedings of Seventh International Conference on Document Analysis and Recognition*, **Vol. 1**: 218–22.

Rath, T. and Manmatha, R. (2003b). *Word Image Matching Using Dynamic Time Warping. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, **Vol. 2,** II–521–7.

Rath, T. M. and Manmatha, R. (2007). Word spotting for historical documents. *IJDAR*, **9**(2–4): 139–52.

Richards, L. G. and Heller, F. P. (1976). Recognition thresholds as a function of word length. *American Journal of Psychology*, **89**(3): 455–66.

Salah, A. B., Ragot, N., and Paquet, T. (2013). Adaptive detection of missed text areas in OCR outputs: Application to the automatic assessment of OCR quality in mass digitization projects. *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, doi:10.1117/12.2003733.

Spitz, A. L. (1997). Determination of the script and language content of document images. *IEEE Transactions Pattern Analysis Machine Intelligence*, **19**(3): 235–45.

Spitz, A. L. (1999). Shape-based word recognition. *ICDAR*, **1**(4): 178–90.

Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond colthearts n: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, **15**(5): 971–9.

Ye, P., Kumar, J., Kang, L., and Doermann, D. S. (2012). *Unsupervised Feature Learning Framework for No-Reference Image Quality Assessment, CVPR*. Providence, RI, USA: IEEE, pp. 1098–105.

Zagoris, K., Papamarkos, N., and Chamzas, C. (2006). *Web Document Image Retrieval System Based on Word Spotting, IEEE International Conference on Image Processing 2006, Atlanta, Georgia, USA*, pp. 477–80.

Zhang, B., Srihari, S. N., and Huang, C. (2004). Word image retrieval using binary features. *Proc. of the SPIE Conf. on Document Recognition and Retrieval XI, 2004, San Jose, California, USA.*