Proceedings of the 2nd International Conference on Computer and Information Science and Technology (CIST'16)
Ottawa, Canada – May 11 – 12, 2016
Paper No. 128

Effect of Typeface Design Features on Legibility and OCR

Jehan Janbi, Ching Y.Suen
Concordia University, CENPARMI
1515 St.Catherine W., Montreal, Canada
jjanbi@gmail.com; suen@cs.concordia.ca

Extended Abstract

Although research on OCR area pursued over decades, very few of them focus on the effect of typeface design on the recognition rate and legibility. CENPARMI conducted two experiments, one for Latin and other for Arabic, aiming to conclude how far OCR recognition rates can be affected by several basic typeface design features. The first experiment measured the OCR accuracy for a set of fonts. For Latin, four OCR tools were used which are ABBY FineReader 6.0 pro, TypeReader 6.0 pro, TextBridge pro millennium, and Omnipage pro 12.0. A text consisting of 2244 words, using a set of 18 Latin typefaces (nine serif and nine sans serif) in 8 and 10 points size, was fed to the four OCRs [1]. For Arabic, a text of 3071 words using a set of 13 Arabic typefaces in sizes 10 and 12 points was fed to Readiris pro 12 Middle East and Sakhr Gold Edition 8 to obtain recognition rate for each type face [2]. Then, a dynamic string matching algorithm (Levenshtein distance) is used to compare OCR outputs with the corresponding input ground truth in order to catch misrecognized letters. The second experiment measured the similarity distances for each pair of letters within the same font to determine which pairs of letters are vulnerable to be misrecognized. For each pair, the Euclidean distance between feature vectors was calculated. Each vector consists of gradient features, magnitude, and directions of the greatest change in intensity in a small neighbourhood for each pixel which were extracted from each character image.

Both scripts have some similar design features, but differ in their meaning depending on the nature of each script. For Latin, the proportions between x-height, ascender and descender are calculated by *XA=xHeight/ascender* and *XD=xHeight/descender*. In addition, the average of vertical and horizontal stem thickness of all letters is calculated as $W_i=T_i/xHeight$ where [i=vertical, horizontal]. Some features like spacing, thickness and serif existence were detected using Fontlab4.5 and human observation. The *xHeight, ascender* and *descender* were determined from the projection profile based on the method used in [3]**.** In Arabic, four different weight proportions *W* were measured by $W_{ij} = T_i/H_i$ where *T* is the stroke thickness of *i=vertical, horizontal* and *H* is the height of *j=loop, tooth* using letters 'Alef' (ا), 'Beh' (ب) and 'Feh' (ف). Another four proportions were calculated using loop and tooth height H_j , average word heights *WH*, loop's and tooth's ascender and descender as $HP_j=H_j/WH$ and $AD_j=ascender_j/descender$ where *j=loop, tooth* [4].

From the result of the two experiments, we have made several observations. In general, a bigger font size provides a better performance in all OCRs with accuracy (93% - 99%) for English and (46% - 91%) for Arabic. For Latin, it was observed that individual letters with serif cause misclassification on (b,h), (u,n), (o,n), (o,u). In addition, fonts with moderate x-height, ascender, and descender fonts had better recognition. That is, if x-height is very short, like with Garamond, central letters, such as (e,c), (o,r), (c,r), (a,s), are misrecognized. Else, if the x-height is larger than ascender and descender, also leads to misrecognition, such as (i,j), (v,y), (g,u), (q,u), (o,g), (f,t) in Heattenschweiler and Impact fonts. Moreover, the extreme and light thickness of stroke decreased the recognition rate because this reduced the inner space of a letter leading to errors like in (e,c), (e,o), (a,e). For Arabic, we observed that the fonts that have loop height around average = 4.88, and tooth height also around average= 5.7 are having high recognition rates. Moreover, the proportion of the ascender to descender should not be very large. Having this proportion around 0.7-0.8 could be adequate to produce a high recognition rate. Also, the letters that have the same basic letter form and differ only in the number of dots, such as 'Sin' (س) and 'Shin' (ش) are frequently misrecognized. In addition to the effect on OCR, results on legibility studies of various typefaces will also be presented [5]. Those observations and results may guide typeface designers to produce more recognizable and legible typefaces.

References

[1] Y. Zhang, “The Effect of Font Design Characteristics on Font Legibility,” MS. Thesis, CENPARMI, Concordia Univ., Montreal, 2006.

[2] M. Saeid, “Discovering the Effect of Arabic Typface Design Characteristics on Font Legibility from OCR Point Of View,” MS. Thesis, CENPARMI, Concordia Univ., Montreal, 2012.

[3] A. Zramdini and R. Ingold, “Optical Font Recognition Using Typographical Features,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 877-882, 1998.

[4] C. Y. Suen, S. Nikfal, B. Zhang, J. Janbi, and N. Dumont, “Personality Traits of Typface for English, Chinese and Arabic,” in *Proceedings of ATypeI Conference*, Hong Kong, 2012.

[5] M. C. Dayson and C. Y. Suen, *Digital Fonts and Reading*. Singapore-London-New Jersey: World Scientific Publishing, 2016.