Ancient Word Indexing Using Fuzzy Methods

Cláudia S. Ribeiro, João M. Gil, João R. Caldas Pinto, João M. Sousa

Technical University of Lisbon, Instituto Superior Técnico, GCAR
Av. Rovisco Pais, 1049-001 Lisboa Portugal

Abstract. This paper proposes an optical word indexing system based on fuzzy logic for 17th century printed documents. A pre-processing stage assures only interesting word candidates are considered. The main image-space indexing procedure builds fuzzy membership functions from oriented features extracted using Gabor filter banks. It proceeds by comparing this data with candidate word features and assigning each a similarity value. Results on a significant test set found over 85% of all word occurrences with a false positive rate of less than 20%.

1 Introduction

This paper proposes an indexing system specifically tailored to ancient documents (17th century) and corresponding typesets based on a handwriting OCR system using fuzzy logic [1]. Indexing is performed from a holistic perspective, i.e., by taking a whole word as a single recognizable symbol, precisely like the original system. The use of fuzzy classification [2] improves results by providing larger tolerance for unstable typesetting and printing technologies. The modifications introduced to [1] are reported and explained along the article. They are mostly related to simplifications inherent to the indexing problem, such as the elimination of word groups and related concepts.

2 Pre-Processing

The pre-processing stage implemented to assist the indexing procedure is based on word image properties, namely aspect ratio filtering.

Intuitively, aspect ratio filtering finds words whose aspect ratio is similar to that of the target word and eliminates the others, not supplying them to the main indexing system. The filtering decision can be expressed mathematically by:

$$f(I) = \left| \frac{ar(I)}{ar(I_T)} - 1 \right| \quad (1)$$

where I is a word image, I_T is the target word image and ar represents an image aspect ratio. If $f(I)$ is greater than a certain positive threshold value, the two aspect ratios are

S. Ribeiro C., M. Gil J., R. Caldas Pinto J. and M. Sousa J. (2004).
Ancient Word Indexing Using Fuzzy Methods.
In *Proceedings of the 4th International Workshop on Pattern Recognition in Information Systems*, pages 210-215
DOI: 10.5220/0002685802100215

deemed too dissimilar and image I is not considered for indexing; otherwise, I is processed by the indexing algorithm.

This filtering procedure was afterwards complemented with a second criterion, based on image area instead of aspect ratio. The filtering decision now additionally considers the following expression:

$$g(I) = \left| \frac{w(I) \times h(I)}{w(I_T) \times h(I_T)} - 1 \right| \quad (2)$$

This function is very similar to f, only image area is compared in place of aspect ratios. This filter is used because, within a given book, equal word images have approximately the same area due to relatively regular character sizes, on average. The final decision value compared with the established threshold is now the maximum between $f(I)$ and $g(I)$, disregarding an image if either filter eliminates it. As shown in the Results section, image area filtering successfully reduces the computational weight of the indexing process without compromising the quality of the results.

3 Fuzzy Indexer

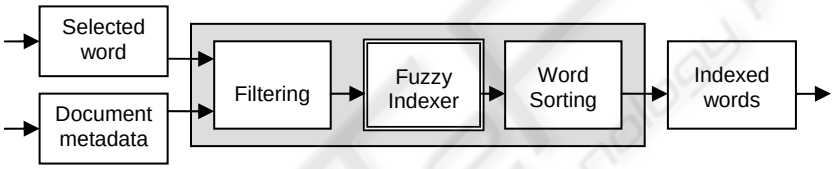

Fig. 1. System Diagram

Fig. 1 displays a diagram representing schematically the organization of the developed system and the streamlined design connecting its components. The input consists in a selected word image and geometrical information that consists in the segmentation coordinates of the searchable words. This information was obtained through the ABBYY FineReader Engine [10].

The main block consists in the fuzzy indexing process. It is described in this section, highlighting particularly the several modifications performed on the original recognizer. Oriented feature extraction will be presented first, then the membership function generation algorithm. Finally, the indexing value computation is explained.

3.1 Feature extraction

The dominant features of a word consist of what is more common not to change between typing styles, as the long vertical stroke in the b's and t's, for example. In this paper their extraction is performed through Gabor filter banks, which allow oriented feature extraction [1]. The original system required a time-consuming alignment algorithm in order to match extracted word features among word group samples. This procedure would make no sense in the current indexing environment,

since there is only one selected word sample, for what could be called a single word group. This simplification reduces the computational effort significantly.

Before the indexing process can truly begin, a set of fuzzy membership functions is generated for each orientation based on the extracted features of the target word image. These intend to provide a description of the image features for use within the indexing algorithm. This process is performed according to [1] except for the changes explained in the next section.

3.2 Indexing

The indexer receives as input a selected target word image and a set of word images where searching is to take place, selected through the image property filtering explained before. The original system considered several word groups, a concept that does not exist in this indexing application. Therefore, most equations had to be adapted to this situation. The more complex modifications are detailed in the remainder of this section, which describes the indexing process. The most important practical difference concerns the absence of a training stage, making the indexing system flexible and lightweight in comparison. The operations corresponding to the training stage are performed swiftly at the beginning of each indexing run.

In the first step of the process, the target word features are extracted and membership functions are generated. Next, each of the candidate images is pre-processed. Namely, each is filtered through Gabor filter banks and resized to match the target image dimensions and enable the remaining calculations. The intensity value of point (x, y) of the resulting image is denoted as $V'_i(x, y)$ for orientation i.

The objective is to compare the image features with the membership function values. A more accurate correspondence should represent a bigger match probability. Points that meet the similarity conditions should be considered, and assigned a positive weight, as well as those that prove dissimilarity, penalizing the rating.

Weights $w_i(x, y)$ were assigned to each image point (x, y), for each orientation i, to measure its influence, related mostly to membership function values. They are calculated according to:

$$w_i(x, y) = V'_i(x, y), \text{ if } a_i(x, y) = 0 \; . \tag{4}$$

$$w_i(x, y) = w'_i(x, y), \text{ if } a_i(x, y) \neq 0 \; . \tag{5}$$

Equation (4) assures that points where V'_i is not zero, but the membership function is, will be penalized. This happens when a feature in V'_i does not match any of orientation i. In this case, the value increases the denominator of Equation (6), while not affecting the numerator, and therefore lowers the computed similarity rating. In [1], a somewhat different and more complex formal notation is used, in part to account for the formal distinction between the various partial membership functions for each word and orientation. In this paper, the partial functions, each corresponding to a particular feature, were iteratively combined, therefore simplifying notation and improving computational resource usage.

Equation (5) simplifies the expression given in [1], where membership function values were combined by weighting their relative importance considering the concept of rate of significance, which refers to the relevance of a given pixel in face of its

intensity values along the various word groups. It is not applicable here, since there are no word groups and no points can be deemed more significant than others for distinguishing among words. Instead, only the membership function value is used, representing the importance of any given pixel as a visual property of the target word.

The original similarity matrix S is now reduced to a vector. Similarity values S_i for each orientation i correspond to an index measuring similarity, within a given orientation, between target and candidate images. Determining these values will consider not only information about the image itself but also about the membership functions and the relative pixel weights. The entries are calculated as follows:

$$S_i = \frac{\sum_{x,y} w_i(x,y) \times a_i(x,y) \times V_i'(x,y)}{\sum_{x,y} w_i(x,y)} \quad (6)$$

These similarity ratings are determined considering the need to penalize the value of points with high $w_i(x, y)$ but low or zero $a_i(x, y)$ or $V'_i(x, y)$. In these cases, the image point has non-zero intensity outside the membership function area or low or zero intensity within the membership function area. This expression is nearly identical to the similarity matrix equation in [1], assuming there is just one word group.

The final indexing stage modifies the simple additive weighted (SAW) method [8] used for recognition. The final indexing value s is computed as follows:

$$s = \sum_{i=1}^{N} v_i \times S_i \text{ , where } v_i = \frac{\sum_{x,y} a_i(x,y)}{\sum_{i,x,y} a_i(x,y)} \quad (7)$$

where N is the number of orientations. The relative membership function weight v_i represents the importance of each orientation in determining the final result. The rationale is that orientations with a relatively larger membership function volume should have a greater influence in the decision-making.

The indexing value expression is based on the word group selection equation, erasing its denominator because it is a mere scale factor in a single-group framework. An indexing value is assigned to each word; it is directly related to the likelihood that the subject word matches the target word. Independently, however, these values have no formal significance. They are useful when compared with each other, so the words are sorted accordingly to achieve practical results.

4 Results

In this section, we present the test results gathered for the purpose of analyzing the software performance and correctness. The test set consists of 20 pages acquired with variable scanning conditions, namely skewing and paper see-through, with both non-italic and italic text. It includes 1886 words segmented by the FineReader engine. The source book [9] concerns Portuguese language orthography, providing a large variety of characters, in the form of word examples during the technical exposition, and significant word repetition, ideal for indexing purposes. Several early tests were performed in order to validate the use of image property filtering. These tests showed that aspect ratio filtering alone eliminated more than 75% of the word candidates.

When coupled with image area filtering, the candidate set is additionally reduced to less than 8% of its original size, without degrading the final results. These tests confirmed the usefulness of the filtering procedure.

Table 1 presents the results obtained with the fuzzy indexer with 20 selected words on the standard test set detailed previously. Analyzing these results leads to several conclusions. Firstly, the base fuzzy indexer was able to find 246 out of 288 words, corresponding to 85.4% success rate. 197 of those (68.4% of all occurrences) were directly at the top of the indexed lists. Although solving the false positive problem is not the primary objective, the results show that it is still within acceptable levels: 46 false positives were detected, less than 19% of all correct matches. Nearly a third were found when processing italic words, a finding justified by the greater visual density of italic typesets. Additionally, thanks to image property filtering, only 152 candidates were considered on average per target word. Therefore, filtering enabled an approximately 12-fold decrease in time usage and reduced memory resource requirements.

Table 1. Indexing results table

Word	Total matches	Candidates (%)	Top matches (%)	False positives (%)	Total found (%)
aberto	33	15.6	39.4	54.5	66.7
Accento	7	7.7	57.1	33.3	85.7
Agudo	5	6.6	80.0	0.0	80.0
contrario	5	5.7	60.0	75.0	80.0
Exceytuaõ-ſe	3	1.5	66.7	33.3	100.0
fechado	30	7.6	26.7	13.0	76.7
meſma	13	15.1	61.5	30.8	100.0
Nas	36	10.3	63.9	7.4	75.0
nomes	25	2.4	96.0	24.0	100.0
Pluraes	11	11.5	100.0	0.0	100.0
primeira	4	6.6	50.0	25.0	100.0
pronunciâ-ſe	4	1.4	100.0	0.0	100.0
que	27	9.8	88.9	4.0	92.6
ſorte	8	11.0	75.0	37.5	100.0
Subſtantivo	5	3.1	40.0	25.0	80.0
tambem	9	10.2	77.8	22.2	100.0
terminações	7	2.4	42.9	50.0	57.1
Verbo	16	15.4	87.5	20.0	93.8
Verbos	32	13.3	84.4	0.0	84.4
Vogal	8	3.7	100.0	0.0	100.0
	Weighted Average	8.0	68.4	18.7	85.4

5 Conclusions

This paper proposed an image-based indexing system based on fuzzy pattern recognition built specifically for 17th century documents. The processing sequence was presented, from the early candidate filtering to the actual computation of similarity values, and test results and procedure were summarized.

The indexer system achieved quality results. Despite some problems detected with compact italic text and small word images with few specific features to extract, most indexing runs returned a large and accurate list of matches, provided word segmentation worked suitably. False matches near the top of the list were limited. The filters developed for indexing performed very well, drastically cutting processing time while retaining high quality output.

Further work can include the development of an automatic parameter adjustment system based on measurable properties of the documents being processed.

Acknowledgements

This work was partly supported by: the “Programa de Financiamento Plurianual de Unidades de I&D (POCTI), do Quadro Comunitário de Apoio III”; the FCT project POSI/SRI/41201/2001; “Programa do FSE-UE, PRODEP III, no âmbito do III Quadro Comunitário de apoio”; and program FEDER. We also wish to express our acknowledgments to the Portuguese Bibioteca Nacional, whose continuous support has made possible this work.

References

1. R. Buse, Z.Q. Liu, J. Bezdek, “Word Recognition Using Fuzzy Logic”, in *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 1, Fev. 2001, pp. 65-76
2. João M.C. Sousa and Uzay Kaymak, "Fuzzy Decision Making in Modeling and Control", World Scientific, Singapore and UK, Dec. 2002
3. R. Buse, Z.Q. Liu, T. Caelli, “A structural and relational approach to handwritten word recognition”, in IEEE *Trans. Syst., Man, Cybern.*, vol. 27, no. 25, Oct. 1997, pp 847-861
4. Parker, J.R., “Algorithms for Image Processing and Computer Vision”, John Wiley & Sons, New York, USA, 1998
5. N. Otsu, "A threshold selection method from gray level histograms", IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, pp. 62-66, 1979
6. Godfried Toussaint, “Solving Geometric Problems with the Rotating Calipers”, in Proc. IEEE MELECON'83, 1983, pp. A10.02/1-4
7. James D. Foley et al, “Computer Graphics – Principles and Practice”, Second Edition in C, Addison-Wesley, Reading, Massachussets, USA, 1990
8. C. L. Hwang, K. Yoon, “Multiple Attribute Decision Making, Methods and Applications, A State-of-the-Art-Survey”, Springer-Verlag, Berlin, Germany, 1981
9. Álvaro Ferreira de Véra, “Orthographia ou modo para escrever certo na lingua Portuguesa”, 17th century, available at Biblioteca Nacional
10. ABBYY FineReader Homepage, http://www.abbyy.com, ABBYY Software House