Efficient Cut-off Threshold Estimation for Word Spotting Applications

A. L. Kesidis[1,2] and B. Gatos[2]

[1]Department of Surveying Engineering
Technological Educational Institution of Athens
GR-12210 Athens, Greece
akesidis@teiath.gr

[2]Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-15310 Athens, Greece
bgat@iit.demokritos.gr

Abstract— **Word spotting is an alternative methodology for document indexing based on spotting words directly on document images with the help of efficient word matching while avoiding conventional OCR procedure. The result of the word spotting procedure is a list of word images ranked according to a certain similarity criterion. In this paper, we propose an efficient method to cut-off the ranked list in order to provide the best tradeoff between recall and precision rates. Our aim is to filter the most relevant results based on a threshold which corresponds to an approximate maximization of the expected F-Measure. This is achieved by introducing an estimator that combines the distance of each ranked word with its cumulative moving average. Experimental results on a database with representative historical printed documents prove the efficiency of the proposed approach.**

Keywords word spotting; cut-off threshold; document indexing

I. INTRODUCTION

Quick and efficient content exploitation is an important feature for any information system that provides access to historical document collections. Such collections usually contain a large number of documents and a robust indexing methodology is an essential performance and efficiency indicator. Due to document degradations, OCR systems often fail to support a correct segmentation of the printed historical documents into individual characters. Word spotting is a content-based retrieval procedure that spots words directly on document images with the help of efficient word matching while avoiding conventional OCR procedures [1], [2]. In the case of historical documents, Rath and Manmatha [3] presented a word matching scheme where noisy handwritten document images are preprocessed into one-dimensional feature sets and compared using the DTW algorithm. Rath et al. [4] present a method for retrieving large collections of handwritten historical documents using statistical models. Lavrenko et al. [5] present a holistic word recognition approach for handwritten historical documents. The query comprises either an actual example from the collection of interest or it is artificially generated from an ASCII keyword. A crucial aspect in the retrieval procedure is the word image representation which relies upon robust features. The retrieval procedure is based on a similarity criterion to be maximized or a distance measure to be minimized [6]. A common approach is to reduce the word representation into a fixed-length vector of features and use geometric distance measures like euclidean, cosine, etc [2][7].

Word spotting produces a list of word images that are ranked according to their distance when compared to the query keyword. An important issue that arises is the proper separation of the ranked results into relevant and irrelevant word instances. This will help to provide the user with only the relevant results in a way that is convenient for searching purposes (e.g. extracting a list of document pages that include one or more instances of the query keyword). Clearly, a fixed similarity threshold cannot be applied since the distribution of distance values obtained may vary when applied to different datasets. In this paper, we propose a method for the determination of a cut-off threshold which is based on an estimator that combines the distance of each ranked word with its cumulative moving average. The efficiency of the proposed method is demonstrated showing that the local maximum of the estimator is highly correlated to the overall maximum of the expected *F*-Measure.

The remainder of this paper is organized as follows: Section II describes the word spotting system. In Section III the proposed methodology is detailed. Section IV presents evaluation results on representative historical documents while in Section V the conclusions are drawn.

II. WORD SPOTTING SYSTEM

The main stages of a word spotting system are (a) word segmentation (b) feature extraction and (c) matching and ranking. In this section we describe how we implemented the main word spotting stages.

A. Word segmentation

The word segmentation of the document pages is performed using the Run Length Smoothing Algorithm (RLSA) [9] which uses dynamic parameters that depend on the average character height as described in [10]. RLSA examines the white runs existing in the horizontal and vertical directions. For each direction, white runs with length less than a threshold are eliminated. The horizontal length threshold is experimentally defined as 50% of the average character height while the vertical length threshold is defined as 10% of the average character height.

B. Feature extraction

The segmented words of the historical document as well as the query keyword are described by feature vectors which

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.64

are used during the matching phase in order to measure similarity between word images. Several features and methods have been proposed in the literature for word image matching based on strokes, contour analysis, etc, [11], [2].

In the proposed approach, two different types of features are combined providing a hybrid features vector for each dataset word as well as for the query keyword [12]. The first one divides the word image into a set of zones and calculates the density of the character pixels in each zone. The second type of features is based on word (upper/lower) profile projections. The word image is divided into two sections with respect to the horizontal line that passes through the center of mass of the word image. Upper/lower word profiles are computed by recording, for each image column, the distance from the upper/lower boundary of the word image to the closest character pixel.

C. Word matching and ranking

The process of word matching involves the comparison/matching between the query keyword image and all the segmented words. Each word image w_i in the document corpus is represented by a feature vector $\boldsymbol{p}_i$, $1 \le i \le N$ in the k-dimensional feature space, where N equals the overall number of words in the dataset. All the words are ranked according to their distance to the k-dimensional feature vector $\boldsymbol{q}$ that represents the input query. The top entries of the ranked list have the smallest distance values and correspond to words that are more similar to the query. As distance metric the cosine similarity is used:

$$d_i = 1 - \frac{\sum_{j=1}^{k} p_{ij} q_j}{\sqrt{\sum_{j=1}^{k} p_{ij}^2 \sum_{j=1}^{k} q_j^2}} \qquad (1)$$

where p_{ij} and q_j are the j-th features of $\boldsymbol{p}_i$ and $\boldsymbol{q}$, respectively.

III. CUT-OFF THRESHOLD DETERMINATION

A. The F-Measure metric

Several methods have been proposed for evaluating the performance of the retrieval system [13], [14]. A well known performance measure is the F-Measure FM which provides a certain tradeoff between specificity and sensitivity. Taking into consideration the top i word instances, FM_i is expressed as the harmonic mean of precision P_i and recall R_i metrics as follows

$$FM(i) = \frac{2 P_i R_i}{P_i + R_i} \qquad (2)$$

Precision P_i is defined as the number of retrieved relevant word instances divided by index i, while recall R_i is defined as the number of relevant word instances divided by the total number of existing relevant words in the dataset. In a typical retrieval scenario, precision is high in the top ranked positions and diminishes gradually while recall follows the reverse direction. F-Measure provides a certain tradeoff between recall and precision with its maximum value indicating the index for which the highest accuracy is achieved.

Fig. 1 demonstrates an example regarding query keyword "famille". The words in the dataset are ranked according to their distance d_i from the query. A subset of the top 50 results is shown in Table I. It can be seen that the top ranking positions are occupied by relevant word instances while irrelevant results start to emerge gradually as the ranking index increases. The maximum value of F-Measure FM_{opt}=0.695 appears for index i_{opt}=48. It is the ranking position that provides the best tradeoff between recall and precision.

Figure 1. Query keyword "famille"

TABLE I. PRECISION, RECALL AND F-MEASURE VALUES FOR SOME OF THE TOP 50 RESULTS OF QEURY KEYWORD "FAMILLE"

Rank i	Word Instance	Distance d_i	Precision P_i	Recall R_i	F-Measure $FM(i)$
1	famille	0.088	1.000	0.021	0.042
2	famille	0.090	1.000	0.043	0.082
3	famille	0.096	1.000	0.064	0.120
…	…	…	…	…	…
21	famille	0.115	0.857	0.383	0.529
22	hostile	0.116	0.818	0.383	0.522
23	famille	0.118	0.826	0.404	0.543
…	…	…			
45	tandis	0.157	0.689	0.660	0.674
46	famille	0.159	0.696	0.681	0.688
47	honnête	0.160	0.681	0.681	0.681
48	famille	**0.160**	**0.688**	**0.702**	**0.695**
49	laquelle	0.161	0.673	0.702	0.688
50	laquelle	0.162	0.660	0.702	0.680
…	…	…	…	…	…

It should be noticed that the actual relevant words are given by the ground truth during the evaluation process and are not known before hand. Our intention is to determine a cut-off threshold index i_{cut} for which $FM(i_{cut})$ is as close as possible to the expected maximum F-Measure score FM_{opt}.

B. *The cut-off threshold*

In order to determine the cut-off threshold which approximately maximizes the expected *F*-Measure we introduce an estimator f_i that combines the distance d_i of the *i*-th ranked word with its cumulative moving average. Let

$$f_i = \frac{d_i}{c_i} \quad (3)$$

where

$$c_i = \frac{1}{i}\sum_{j=1}^{i} d_j \quad (4)$$

is the cumulative moving average. For $i>1$ the *i*-th value of c_i can be calculated recursively by the current d_i value and the previous c_{i-1} as follows

$$c_i = \frac{d_i + (i-1)c_{i-1}}{i} \quad (5)$$

Fig. 2 depicts d_i, c_i and f_i for all the *N* ranked words in the dataset. It can be seen that the distance values d_i grow rapidly up to a value (~0.2) and then there is a large amount of words whose distance from the query is between 0.2 and 0.4. For the last ranked words the distance is exponentially increasing again. The values of f_i reach an overall maximum for $i=N$ that equals

$$\|f\|_\infty = f_N = \frac{N d_N}{c_N} \quad (6)$$

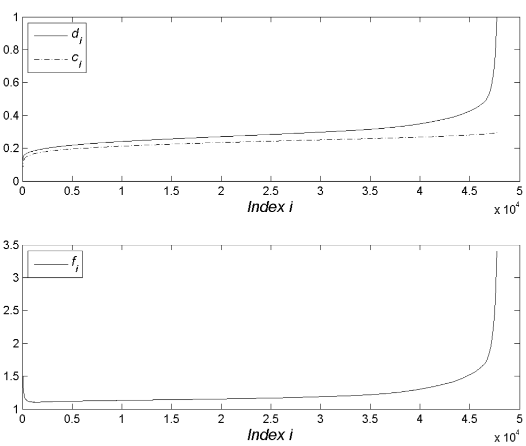


Figure 2. Values of d_i, c_i and f_i for all the words in the dataset.

Besides its overall maximum, f_i reaches a local maximum for relatively small values if index *i*. This can be seen more clearly in the example of Fig. 3 which focuses on the vicinity of the top ranked words. Both d_i and c_i are monotonically non-decreasing curves and the local maximum of f_i appears when their ratio is maximized, as shown in the bottom subplot of Fig. 3.

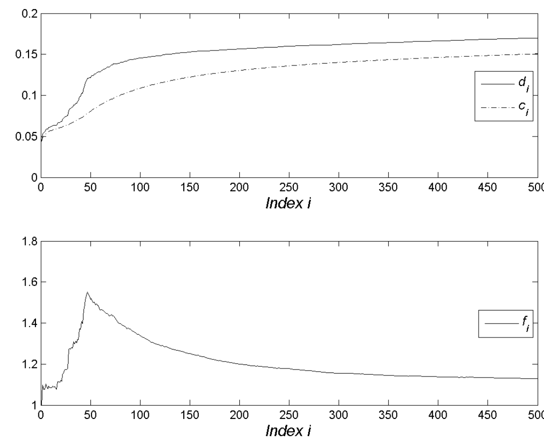


Figure 3. Values of d_i, c_i and f_i for the top r=500 ranked words.

We have noticed that setting the cut-off threshold i_{cut} equal to the index that locally maximizes f_i provides a good estimation of the expected maximum *F*-Measure, that is

$$i_{cut} = \underset{i\in[1...r]}{\arg\max}(f_i) \quad (7)$$

where f_i is considered in the vicinity of the top *r* words. Index *r* is determined by the ranking position that corresponds to the word whose distance is closest to the mean word distance. As mean word distance d_m we denote the cosine distance of the query vector $\boldsymbol{q}$ from a vector $\boldsymbol{m}$ whose *j*-th element equals the mean *j*-th feature of all the word feature vectors. That is,

$$r = \underset{i}{\arg\min}(|d_i - d_m|) \quad (8)$$

where

$$d_m = 1 - \frac{\sum_{j=1}^{k} m_j q_j}{\sqrt{\sum_{j=1}^{k} m_j^2 \sum_{j=1}^{k} q_j^2}} \quad (9)$$

and

$$\boldsymbol{m} = \left(\bar{p}_1, \bar{p}_2, ..., \bar{p}_k\right) \quad (10)$$

Fig. 4 depicts an example where the maximum of *F*-Measure appears at index i_{opt}=argmax(FM_{opt})=43 and equals FM_{opt}=0.805 while i_{cut}=47 and $FM(i_{cut})$=0.79. Even when indexes i_{cut} and i_{opt} differ significantly, as shown in the example of Fig. 5 the corresponding *F*-Measure values are highly correlated.

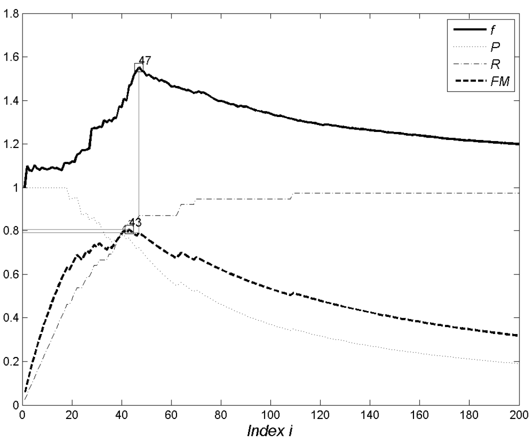


Figure 4. The maximum of *F*-Measure is at index i_{opt}=43 and equals FM_{opt}=0.805. The cut-off threshold is i_{cut}=47 and corresponds to *F*-Measure $FM(i_{cut})$=0.79.

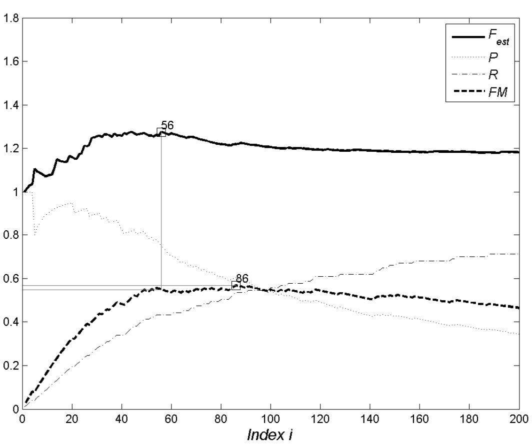
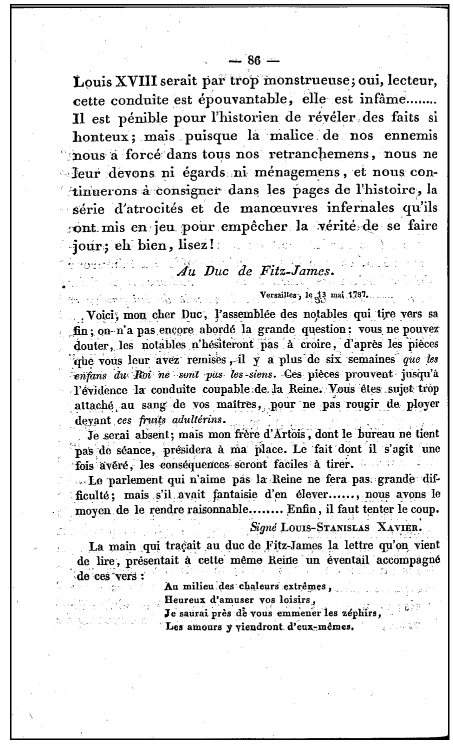


Figure 5. The maximum of *F*-Measure is at index i_{opt}=86 and equals FM_{opt}=0.568. The cut-off threshold is i_{cut}=56 and corresponds to *F*-Measure $FM(i_{cut})$=0.549, very close to FM_{opt}.

IV. EXPERIMENTAL RESULTS

We tested our methodology on a French historical book which was published in 1838 and is owned by Bibliothèque Nationale de France. In Fig. 6 a sample image is shown. We selected 153 pages from this book that contain an overall of 47715 words. We manually marked the ground truth for 20 keywords that have a variety of instances ranging from 33 up to 362, as shown in Table II.



— 86 —

Louis XVIII serait par trop monstrueuse; oui, lecteur, cette conduite est épouvantable, elle est infâme........ Il est pénible pour l'historien de révéler des faits si honteux; mais puisque la malice de nos ennemis nous a forcé dans tous nos retranchemens, nous ne leur devons ni égards ni ménagemens, et nous continuerons à consigner dans les pages de l'histoire, la série d'atrocités et de manœuvres infernales qu'ils ont mis en jeu pour empêcher la vérité de se faire jour; eh bien, lisez!

Au Duc de Fitz-James.

Versailles, le 13 mai 1787.

Voici, mon cher Duc, l'assemblée des notables qui tire vers sa fin; on n'a pas encore abordé la grande question: vous ne pouvez douter, les notables n'hésiteront pas à croire, d'après les pièces que vous leur avez remises, il y a plus de six semaines *que les enfans du Roi ne sont pas les siens*. Ces pièces prouvent jusqu'à l'évidence la conduite coupable de la Reine. Vous êtes sujet trop attaché au sang de vos maîtres, pour ne pas rougir de ployer devant *ces fruits adultérins*.

Je serai absent; mais mon frère d'Artois, dont le bureau ne tient pas de séance, présidera à ma place. Le fait dont il s'agit une fois avéré, les conséquences seront faciles à tirer.

Le parlement qui n'aime pas la Reine ne fera pas grande difficulté; mais s'il avait fantaisie d'en élever......, nous avons le moyen de le rendre raisonnable........ Enfin, il faut tenter le coup.

Signé LOUIS-STANISLAS XAVIER.

La main qui traçait au duc de Fitz-James la lettre qu'on vient de lire, présentait à cette même Reine un éventail accompagné de ces vers:

Au milieu des chaleurs extrêmes,
Heureux d'amuser vos loisirs,
Je saurai près de vous emmener les zéphirs,
Les amours y viendront d'eux-mêmes.

Figure 6. A document image sample

TABLE II. QUERY KEYWORDS AND THEIR INSTANCES

Keyword	Instances	Keyword	Instances
avait	121	jamais	55
bien	97	justice	44
comme	102	Louis	156
contre	52	Madame	39
dans	325	mort	51
fait	104	plus	196
famille	47	pour	362
France	44	Roi	56
homme	39	sans	97
hommes	33	tous	102

For each query keywords the precision *P*, recall *R*, and *F*-Measure curves are calculated according to the ground truth. Based on these measures, the maximum value FM_{opt} of *F*-Measure and its index i_{opt} are determined. The index i_{cut} of f_i is given by (7) and the corresponding *F*-Measure value $FM(i_{cut})$ is also calculated. The results for all the query keywords are shown in Table III. The last column presents the efficiency of FM_{opt} approximation. The experimental results show that the proposed estimator *f* shows a consistency in approximating the FM_{opt} value. Moreover, *f* does not depend on the ranking position of index i_{cut}. Indeed, there are cases where i_{cut} and i_{opt} are nearby indexes and $FM(i_{cut})$ provides an almost perfect approximation of FM_{opt} regardless if i_{cut} points to a small cut-off threshold (e.g. i_{cut}=44 for “contre”) or it indicates a much higher threshold (e.g. i_{cut}=140 for “Louis”). Even more interesting are cases where despite that indexes i_{cut} and i_{opt} differ significantly (e.g. “bien” and “tous”), $FM(i_{cut})$ still provides a good approximation to the maximum *F*-Measure value. Even in case of keyword “avait”, where index i_{cut} is almost two times larger than i_{opt}, the estimation error remains low, i.e. about 11% of FM_{opt}. The overall experimental results show that the average estimation error is below 8% in terms of *F*-Measure.

TABLE III. APPROXIMATION EFFICIENCY FOR SEVERAL CUT OFF THRESHOLDS

Query	i_{opt}	FM_{opt}	i_{cut}	$FM(i_{cut})$	$FM(i_{cut})/FM_{opt}$*100
avait	85	0.495	160	0.441	89.12%
bien	86	0.568	56	0.549	96.61%
comme	98	0.770	100	0.762	99.01%
contre	43	0.863	44	0.854	98.96%
dans	272	0.807	403	0.709	87.79%
fait	69	0.358	240	0.291	81.11%
famille	48	0.695	46	0.688	99.06%
France	19	0.413	25	0.377	91.30%
homme	17	0.500	12	0.431	86.27%
hommes	27	0.700	16	0.612	87.46%
jamais	53	0.519	43	0.469	90.52%
justice	31	0.480	45	0.449	93.63%
Louis	136	0.897	140	0.892	99.40%
Madame	43	0.805	47	0.791	98.24%
mort	34	0.706	38	0.697	98.69%
plus	140	0.821	157	0.805	97.94%
pour	381	0.770	515	0.673	87.39%
Roi	51	0.748	53	0.734	98.17%
sans	68	0.727	187	0.570	78.43%
tous	144	0.528	115	0.525	99.41%
				Total average	**92.93%**

V. CONCLUSIONS

This paper proposes an efficient method for the estimation of a cut-off threshold that can be applied to the ranked results list of a word spotting system in order to filter the most relevant words. As a performance measure, the *F*-Measure is used which provides a certain tradeoff between specificity and sensitivity. The method is based on an estimator that for each ranked word combines its distance with its cumulative moving average. The estimator has a local maximum which is highly correlated to the overall maximum of the expected *F*-Measure. Experiments on a database with representative historical printed documents evidenced promising results that demonstrate the feasibility of the proposed method.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT).

REFERENCES

[1] A. Murugappan, B. Ramachandran, P. Dhavachelvan, “A survey of keyword spotting techniques for printed document images”, Artificial Intelligence Review, Vol. 35, No 2, pp. 119-136, 2011.

[2] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis and S. J. Perantonis, "Keyword-Guided Word Spotting in Historical Printed Documents Using Synthetic Data and User feedback", International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, Vol. 9, No. 2-4, pp. 167-177, 2007.

[3] T. M. Rath, and R. Manmatha, “Features for word spotting in historical documents” In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03) , pp 218-222, 2003.

[4] T. M. Rath, R. Manmatha and V. Lavrenko “A search engine for historical manuscript images” ACMSIGIR Conference, pp. 369–376, 2004.

[5] V. Lavrenko, T. M. Rath and R. Manmatha, “Holistic word recognition for handwritten historical documents”, Proceedings of the International Workshop on Document Image Analysis for Libraries, pp. 278–287, 2004.

[6] R. Silva, R. Stasiu, V. M. Orengo and C. A. Heuser, “Measuring quality of similarity functions in approximate data matching”, Journal of Informatics, pp 35-46, 2007.

[7] A. Kesidis, E. Galiotou, B. Gatos, A. Lampropoulos, I. Pratikakis, I. Manolessou and A. Ralli, “Accessing the content of greek historical documents”, Proc. of Workshop on Analytics for Noisy Unstructured Text Data, (AND ’09), pp. 55–62, 2009.

[8] A. Arampatzis, J. Kamps, and S. Robertson, “Where to Stop Reading a Ranked List? Threshold Optimization using Truncated Score Distributions”, Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 19-23, 2009.

[9] F.M. Wahl, K. Y. Wong and R. G. Casey, "Block segmentation and text extraction in mixed text/image documents", Comput. Graph. Image Process. Vol. 20, pp. 375–390, 1982.

[10] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos and N. Papamarkos, "Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths", Image and Vision Computing, Vol. 28 , Issue 4, pp. 590-604, 2010.

[11] D. Doerman, H. Li, O. Kia “The detection of duplicates in document image databases”, Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR’97), pp. 314–318, 1997.

[12] A. L. Kesidis, E. Galiotou, B. Gatos and I. Pratikakis, “A word spotting framework for historical machine-printed documents”, International Journal on Document Analysis and Recognition, DOI: 10.1007/s10032-010-0134-4, pp. 1-14, 2010.

[13] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[14] H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, A.E. Abbadi, “Approximate Nearest Neighbor Searching in Multimedia Databases”, Proc. of 17th International Conference on Data Engineering, (ICDE ’01), pp. 503-511, 2001.