ICFHR 2014 Competition on Handwritten KeyWord Spotting (H-KWS 2014)

Ioannis Pratikakis, Konstantinos Zagoris
Visual Computing Group
Department of Electrical and Computer Engineering
Democritus University of Thrace
Xanthi, Greece
Email: {ipratika, kzagoris}@ee.duth.gr

Basilis Gatos, Georgios Louloudis and Nikolaos Stamatopoulos
Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos"
Athens, Greece.
Email: {bgat, louloud, nstam}@iit.demokritos.gr

***Abstract*—H-KWS 2014 is the Handwritten Keyword Spotting Competition organized in conjunction with ICFHR 2014 conference. The main objective of the competition is to record current advances in keyword spotting algorithms using established performance evaluation measures frequently encountered in the information retrieval literature. The competition comprises two distinct tracks, namely, a segmentation-based and a segmentation-free track. Five (5) distinct research groups have participated in the competition with three (3) methods for the segmentation-based track and four (4) methods for the segmentation-free track. The benchmarking datasets that were used in the contest contain both historical and modern documents from multiple writers. In this paper, the contest details are reported including the evaluation measures and the performance of the submitted methods along with a short description of each method.**

Keywords*—Word Spotting, Handwritten Documents, Benchmarking

I. Introduction

Handwritten keyword spotting is the task of detecting query words in handwritten document image collections without involving a traditional OCR step. Recently, handwritten word spotting has attracted the attention of the research community in the field of document image analysis and recognition since it appears to be a feasible solution for indexing and retrieval of handwritten documents in the case that OCR-based methods fail to deliver satisfactory results. In the Handwritten KeyWord Spotting 2014 (H-KWS 2014) competition, an evaluation framework is established for benchmarking handwritten keyword spotting approaches which address the query by example problem. The task considered for evaluation is as follows. A query word image along with a collection of handwritten document images is provided as input to the system under evaluation. The expected output is a ranked list of bounding boxes, which correspond to spotted word images that match the query word image in terms of a similarity value.

The competition has two distinct tracks for handwritten keyword spotting: TRACK I - Segmentation-based and TRACK II - Segmentation-free. For TRACK I, the location of the word images in the document images of the dataset is given.

The objective of the H-KWS 2014 is threefold:

- Record current advances in keyword spotting.
- Provide benchmarking handwritten datasets[1] containing both historical and modern documents from multiple writers.
- Explore established evaluation performance measures frequently encountered in the information retrieval literature while providing the software for these measures as implementation reference[2].

The remainder of the paper is structured as follows: Section II provides the description of the methodology used in the competition by each participant. Section III presents the competition datasets. The evaluation measures are detailed in Section IV while the discussion of the competition results is given in Section V.

II. Methods and Participants

Five (5) distinct research groups have participated in the competition with three (3) methods for the segmentation-based track and four (4) methods for the segmentation-free track. Brief description for each method is given in the following (the order of appearance reflects the chronological order of expressing an interest to participate in the competition).

1. The Blavatnik School of Computer Science, Tel-Aviv University, Israel *(Alon Kovalchuk, Lior Wolf, Nachum Dershowitz) (TRACK I, TRACK II)*: The dataset images that are to be queried are preprocessed by a simple binarization operation, followed by the extraction of multiple overlapping candidate targets. The number of candidates is 30 (2 times the number of actual words depending page quality). The latter stage is unnecessary while evaluating segmented pages. Each binary target, as well as the binarized query, is resized to fit a fixed-size rectangle and represented by conventional image descriptors. The resized image of size 160x56 is divided into a grid of 20x7 cells, each of 8x8 pixels. Using HOG ($\Re^{31}$) and LBP ($\Re^{58}$) descriptors for each cell, we get a vector of size 31x20x7+58x20x7 = 12,460. Then, a cosine similarity operator – followed by maximum pooling over random groups – is used to represent each target or query as a concise 250D vector. To improve query results we move query image to each of 4 directions and do max pooling on the results. Top results can be re-evaluated using also the original vector in addition to

[1] http://vc.ee.duth.gr/h-kws2014/#Datasets
[2] http://vc.ee.duth.gr/h-kws2014/#VCGEval

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.142

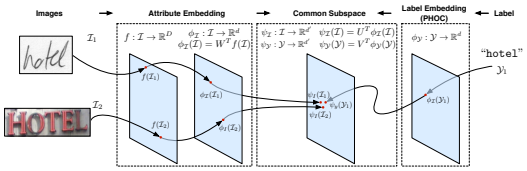
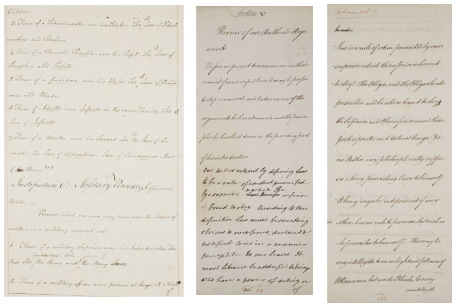


Fig. 1. Overview of the Almazán method.

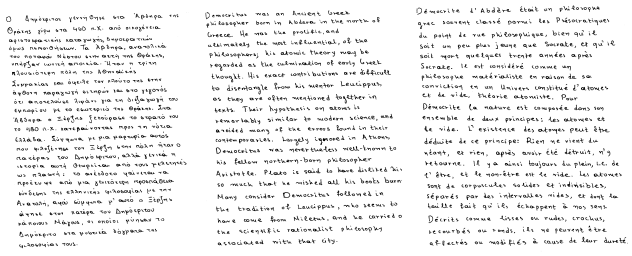
250D vector. Retrieval is performed in a fraction of a second by nearest-neighbor search within that space, followed by a simple suppression of extra overlapping candidates.

2. Computer Vision Center, Universitat Autònoma de Barcelona, Spain *(Jon Almazán, Albert Gordo, Ernest Valveny) (TRACK I)*: The methodology is based on the work that was proposed in [1], where the spotting and recognition tasks were addressed by learning a common representation for word images and text strings. In this work, character attributes are used to learn a semantic representation of the word images and then perform a calibration of the scores with CCA that puts images and text strings in a common subspace. After that, spotting and recognition become simple nearest neighbor problems in a very low dimensional space. This method consists of the following process: First, text strings are embedded into a $d-$dimensional binary space – dubbed pyramidal histogram of characters or PHOC – that encodes if a particular character appears in a particular spatial region of the string.
Then, this embedding is used as a source of character attributes: we will project word images into another $d-$dimensional space, more discriminative, where each dimension encodes how likely that word image contains a particular character in a particular region, in obvious parallelism with the PHOC descriptor.
However, due to some differences, direct comparison is not optimal and some calibration is needed. Finally a low-dimensional common subspace is learned with an associated metric between the PHOC embedding and the attributes embedding. The advantages of this are twofold. First, it makes direct comparison between word images and text strings meaningful. Second, attribute scores of images of the same word are brought together since they are guided by their shared PHOC representation. An overview of the method can be seen in Fig. 1. A Matlab implementation of the code can be found in [2]. Since the writing styles of both Bentham and Modern dataset are quite different, different character attributes have been used that have been learned in different training data. For the Bentham dataset, attributes have been learned in the George Washington, and for the Modern dataset the IAM dataset was used.

3. Smith College, Department of Computer Science, Northampton MA, USA *(Nicholas R. Howe) (TRACK I, TRACK II)*: Both the segmentation-based and segmentation-free entries employ the flexible template mechanism described in ICDAR 2013 paper [3]. In both cases, a flexible inkball model is derived from the query image (after binarization if necessary). This is a generative model for word appearance, and allows for Gaussian random-walk deformation of the ink trace in two dimensions. Query models are fit to the target page images to find locations where there is a good (low-

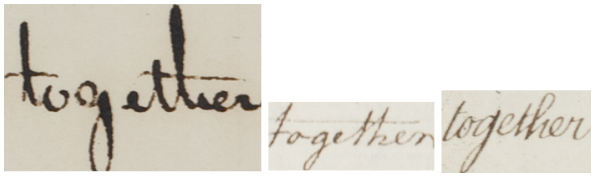


(a)

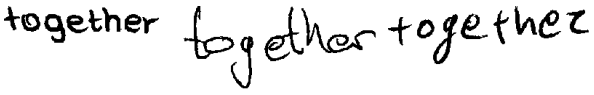


(b)

Fig. 2. Example document images from (a) Bentham Dataset, (b) Modern Dataset



(a)



(b)

Fig. 3. Representative cases showing the variations of the word "together" in (a) Bentham dataset, (b) Modern dataset

deformation) match. The best candidate match regions are themselves converted to flexible inkball models and fit back to the query image for reverse verification. Regions with the best two-way match scores are ranked highest in the returned list of hit locations.
The entries in the two contest tracks differ slightly in the details of their implementation. The segmentation-free algorithm matches the query model to full target pages without any adjustments for scale. Although theoretically beneficial, attempting to match scales between the query and target text introduces the potential for error if the scale estimation is incorrect. The reverse matching step uses the ink in the immediate neighborhood of the query model match, without attempting any word segmentation, and may therefore allow matching on a word fragment. In contrast, the segmentation-based algorithm attempts to scale the query model to each target word by matching the interquartile distance of both the horizontal and vertical ink projections, limited by a few heuristics to prevent extreme stretching. Since the target word

bounding boxes are known in the segmentation-based contest, only full words are used in the reverse matching step.

4. Université de Lyon, CNRS, INSA-Lyon, LIRIS, France *(Yann Leydier, Frank Lebourgeois) (TRACK II)*: This method [4], [5] is learning-free and segmentation-free and can be applied directly on colour and greyscale images without binarisation. It has been successfully tested on most of the kinds of manuscripts: medieval Latin, Arabic, Chinese, Sanskrit, and even hieroglyphs and cuneiform. Its drawback is that the method is rather slow and is generally writer dependant.
Zones of interest are extracted from the query-word. These zones are composed of high-curvature locations that correspond to the strokes' extrema and intersections. The geometric links between the zones of interest are elastic, so that the query-word can be deformed cohesively to fit the variability of the manuscripts.
The query-graph of zones of interest is compared to the image on multiple locations corresponding to strokes. For each zone of interest, the gradient angles of the query and the image are compared. Unlike many recent works, histograms of gradients (HOGs) are not used, thus our method is slower but much more accurate. The score is not a real distance but a kind of inclusion measure. The results are sorted by ascending score.
In its general from, there are three query options: word-spotting (classical query-by-example), sketch-spotting (the query is hand-drawn by the end-user) and word retrieval (the end-user types a plain text query-word and the query-graph of zones of interest is composed from a list of images of characters that were manually extracted from the document).

5. Institute for Communications Technology (IfN) of Technische Universität Braunschweig, Braunschweig, Germany *(Werner Pantke, Martin Dennhardt, Volker Märgner, Tim Fingscheidt) (TRACK II)*: For the competition, a template-based word spotting approach was used, previously presented in [6], which is derived from [4] and [5]. Being developed during the course of the HADARA project [7], the typical task of this spotting is to find word occurrences of a given keyword template in large historical documents, which are rarely written by the same writer [8], but typically written in the same writing style. Paleographers, which have to take a deep look into writing styles, may use this system to search for words written in a specific significant writing style. The spotting method operates on images that are not segmented containing color or gray value information.
The methodology can be outlined as follows. First, a shading correction is applied and, in the case of color, the images are converted to grayscale using a pseudo luminance approach. Gradient angle and magnitude are exploited as features, which are compared within automatically identified *zones of interest* (ZOIs) using a cohesive elastic matching [4]. This matching technique is robust against typical variations that can be found in handwritten text, but is not able to detect completely different pen and writing styles. ZOIs are found by locating local maxima of the gradient field curvature of the template image. To reduce the computational costs, the template is only matched against document areas assumed to contain words. Output of the algorithm is an n-best list of areas in the given set of document images that resemble the shape of the template image, with n denoting the number of results. In compliance with [9], multiple hits of the same word occurrence are tried to be avoided by filtering out overlapping results from the result list. For more details of this word spotting approach, you should refer to work in [6].
For this competition, the following changes were made to the submitted word spotting methodology. Due to the limited competition time, the amount that each possible pair of template ZOIs is allowed to overlap is lowered, resulting in less ZOIs per template image and, thus, decreasing the computational effort with the drawback of a lower accuracy. As the Modern set of the competition consists of binary images without grayscale or color information, it was necessary for the feature extraction to re-obtain grayscale images. For this purpose, a Gaussian filter is applied to each image before any processing.

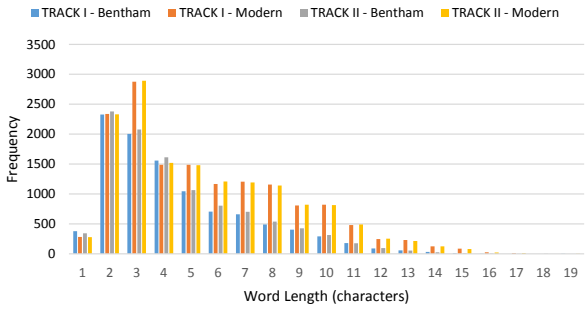
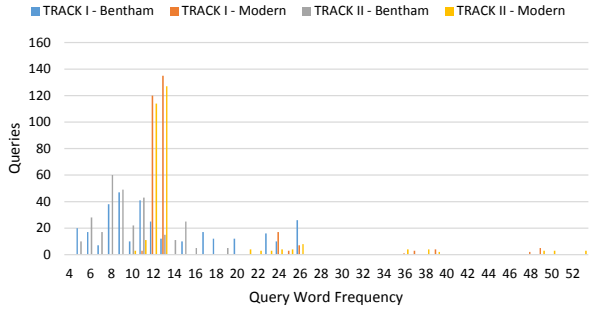
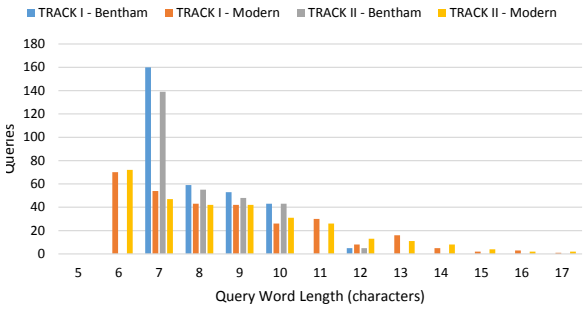


Fig. 4. Dataset word length frequencies



(a)



(b)

Fig. 5. Statistics for: (a) the frequency of the query words in each dataset, (b) the length of the queries

III. Datasets and Query Sets

As described in Section I the competition comprises two tracks which differ on whether they encounter word im-

TABLE I. RELEVANCE JUDGEMENTS FOR THE QUERY WORD "*husband*"

Word	Relevance Judgement
husband	1.0
husband	1.0
husband	1.0
husband	1.0
husband	1.0
husband,	0.9
husband:	0.9
husband.	0.9
Husband.	0.8
Husband]	0.8

age segmentation in a document (segmentation-based) or not (segmentation-free). The competition for both tracks considers the following datasets:

Bentham Dataset [10]: It consists of high quality (approximately 3000 pixels width and 4000 pixels height) handwritten manuscripts. The documents are written by Jeremy Bentham (1748-1832) himself as well as by Bentham's secretarial staff over a period of sixty years.

Modern Dataset: It consists of modern handwritten documents from the ICDAR 2009 Handwritten Segmentation Contest [11]. These documents originate from several writers that were asked to copy a given text. They do not include any non-text elements (lines, drawings, etc.) and are written in four (4) languages (English, French, German and Greek).

For each track, 50 document images of Bentham dataset and 100 document images of Modern dataset (25 documents per language) were used for testing at the competition, resulting in a total of 300 document images for both tracks.

Fig. 2 shows some representative document images from these datasets. They both contain several very difficult problems to be addressed, wherein the most difficult is the word variability. The variation of the same word is high and involves writing style, font size, noise as well as their combination. Fig. 3 shows the variations that may appear for a particular word. The word-length statistics for each dataset are shown in Fig. 4.

The query set of each dataset is provided in XML format and it contains word image queries of length greater than 6 and frequency greater than 5. Fig. 5a depicts the frequency of each query set in the dataset while Fig. 5b presents word length statistics for each query set.

Both, datasets and query sets, can be downloaded from http://vc.ee.duth.gr/h-kws2014/#Datasets.

IV. EVALUATION MEASURES

The measures employed in the performance evaluation of the submitted word spotting algorithms are the Precision at Top 5 Retrieved words (P@5), the Mean Average Precision (MAP) and the Normalized Discounted Cumulative Gain (NDCG) for both binary and non-binary judgement relevancies. Finally, Precision-Recall Curves are provided to showcase the methods performance across the recall range.

To further detail the metrics, let define Precision and P@k as follows:

$$P@k = \frac{|\{relevant\ words\} \cap \{k\ retrieved\ words\}|}{|\{k\ retrieved\ words\}|} \quad (1)$$

Precision is the fraction of retrieved words that are relevant to the query, while in the case that precision should be determined for the k top retrieved words, P@k is computed. In particular, in the proposed evaluation, P@5 is used which is the precision at top 5 retrieved words. This metric defines how successfully the algorithms produce relevant results to the first 5 positions of the ranking list.

The second metric used in the proposed evaluation is the Mean Average Precision (MAP) which is a typical measure for the performance of information retrieval systems [12], [13]. It is implemented from the Text REtrieval Conference (TREC) community by the National Institute of Standards and Technology (NIST). The above metric is defined as the average of the precision value obtained after each relevant word is retrieved:

$$AP = \frac{\sum_{k=1}^{n} (P@k \times rel(k))}{\{relevant\ words\}} \quad (2)$$

where:

$$rel(k) = \begin{cases} 1, & if\ word\ at\ rank\ k\ is\ relevant \\ 0, & if\ word\ at\ rank\ k\ is\ not\ relevant \end{cases} \quad (3)$$

In the competition context, non-binary relevance judgement is introduced by incorporating the Normalized Discounted Cumulative Gain (NDCG) metric in order to deal with small variations of the query word that can be found in the datasets. Table I shows an example of those non-binary relevance judgement for the query word "husband".

The NDCG measures the performance of a retrieval system based on the graded relevance of the retrieved entities. It varies from 0.0 to 1.0, with 1.0 representing the ideal ranking of the entities.

It is defined as:

$$nDCG = \frac{DCG}{IDCG} \quad (4)$$

where:

$$DCG = rel_1 + \sum_{i=1}^{n} \frac{rel_i}{log_2(i+1)} \quad (5)$$

where rel_i is the relevance judgement at position i, and $IDCG$ is the ideal DCG which is computed from the perfect retrieval result.

Contrary to the non-binary Relevance Judgement values, for the binary NDCG, the value '1' is employed.

The Precision - Recall Curve is calculated by the traditional 11-point interpolated average precision approach [14], [15]. For each query, the interpolated precision is measured at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0.

Additionally, segmentation - free systems impose supplementary problems as they may not detect the whole word

TABLE II. EXPERIMENTAL RESULTS FOR TRACK I: SEGMENTATION-BASED.

	BENTHAM DATASET				***MODERN DATASET***			
Method	**P@5**	**MAP**	**NDCG (Binary)**	**NDCG**	**P@5**	**MAP**	**NDCG (Binary)**	**NDCG**
G1	0.738 (1)	0.524 (1)	0.742 (2)	0.762 (2)	0.588 (2)	0.338 (2)	0.611 (2)	0.612 (2)
G2	0.724 (2)	0.513 (2)	0.744 (1)	0.764 (1)	0.706 (1)	0.523 (1)	0.757 (1)	0.757 (1)
G3	0.718 (3)	0.462 (3)	0.638 (3)	0.657 (3)	0.569 (3)	0.278 (3)	0.484 (3)	0.485 (3)

TABLE III. EXPERIMENTAL RESULTS FOR TRACK II: SEGMENTATION-FREE

BENTHAM DATASET																
	P@5				**MAP**				**NDCG (Binary)**				**NDCG**			
Method	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average
	0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8	
G1	0.617	0.611	0.599	0.609 (1)	0.428	0.419	0.402	0.416 (1)	0.653	0.640	0.621	0.638 (1)	0.671	0.657	0.640	0.56 (1)
G3	0.596	0.568	0.506	0.556 (2)	0.397	0.372	0.321	0.363 (2)	0.551	0.518	0.457	0.509 (2)	0.569	0.536	0.474	0.526 (2)
G4	0.351	0.341	0.313	0.335 (4)	0.219	0.209	0.187	0.205 (4)	0.386	0.363	0.319	0.356 (4)	0.400	0.376	0.331	0.369 (4)
G5	0.597	0.55	0.477	0.543 (3)	0.385	0.347	0.280	0.337 (3)	0.569	0.513	0.424	0.502 (3)	0.586	0.531	0.440	0.519 (3)
MODERN DATASET																
	P@5				**MAP**				**NDCG (Binary)**				**NDCG**			
Method	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average
	0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8	
G1	0.541	0.541	0.535	0.539 (1)	0.265	0.265	0.259	0.263 (1)	0.491	0.484	0.473	0.483 (1)	0.491	0.485	0.474	0.483 (1)
G3	0.429	0.422	0.399	0.417 (2)	0.170	0.165	0.152	0.163 (2)	0.310	0.301	0.277	0.296 (2)	0.310	0.301	0.277	0.296 (2)
G4	0.250	0.241	0.211	0.234 (4)	0.095	0.089	0.077	0.087 (4)	0.218	0.195	0.161	0.191 (4)	0.218	0.195	0.161	0.191 (4)
G5	0.264	0.247	0.223	0.245 (3)	0.100	0.092	0.081	0.091 (3)	0.229	0.201	0.168	0.199 (3)	0.229	0.202	0.168	0.200 (3)

or they include parts of another word. A word instance is considered as detected only if there is a significant overlap with the ground truth word. The overlap is expressed by the intersection over the ground truth word area metric (IOA) and it is defined as: $IOA = \frac{A \cap B}{A}$, where A and B denote the bounding box areas of the ground truth word and the method output word, respectively. The IOA metric ranges from 0 to 1, where 1 corresponds to exact matching. A threshold T is applied in order to decide whether the word instance and the segmented word match sufficiently. In this case, the performance evaluation for three different thresholds (0.6, 0.7 and 0.8) is used for testing.

Moreover, an evaluation application is developed as referenced implementation for each metric. It is available for Windows, Mac OS X and Linux operating systems as both command-line and GUI form. It accepts as input the experimental results file and the relevance judgement file, which represents the ground truth. Afterwards, it calculates the aforementioned evaluation metrics. The program can be downloaded at http://vc.ee.duth.gr/h-kws2014/#VCGEval and the competition ground truth at http://vc.ee.duth.gr/h-kws2014/#Resources

V. EVALUATION RESULTS

For the sake of clarity, it should be noted that each algorithm appears with the enumeration of the group as presented at Section II. For example, a method submitted by the group No. 3 will appear as G3 (Group 3).

Tables II and III show the evaluation results of each competing algorithm while Fig. 6 shows the Precision - Recall Curves. Inside the parenthesis is the ranking value between the competing methods for the corresponding algorithm. The summation of all accumulated ranking values for all evaluation metrics denote the final score which is shown in Table IV. Overall, for TRACK I the best performance is achieved by **Method G2** which has been submitted by **Jon Almazán, Albert Gordo, Ernest Valveny** affiliated to the **Computer Vision Center, Universitat Autònoma de Barcelona, Spain**. For TRACK II, the best performance is achieved by **Method G1** which has been submitted by **Alon Kovalchuk, Lior Wolf, Nachum Dershowitz** affiliated to the **The Blavatnik School of Computer Science, Tel-Aviv University, Israel**.

TABLE IV. FINAL RANKING LIST

TRACK I: SEGMENTATION-BASED		
Rank	Method	Score
1	G2	10
2	G1	14
3	G3	24
TRACK II: SEGMENTATION-FREE		
Rank	Method	Score
1	G1	8
2	G3	16
3	G5	24
4	G4	32

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme

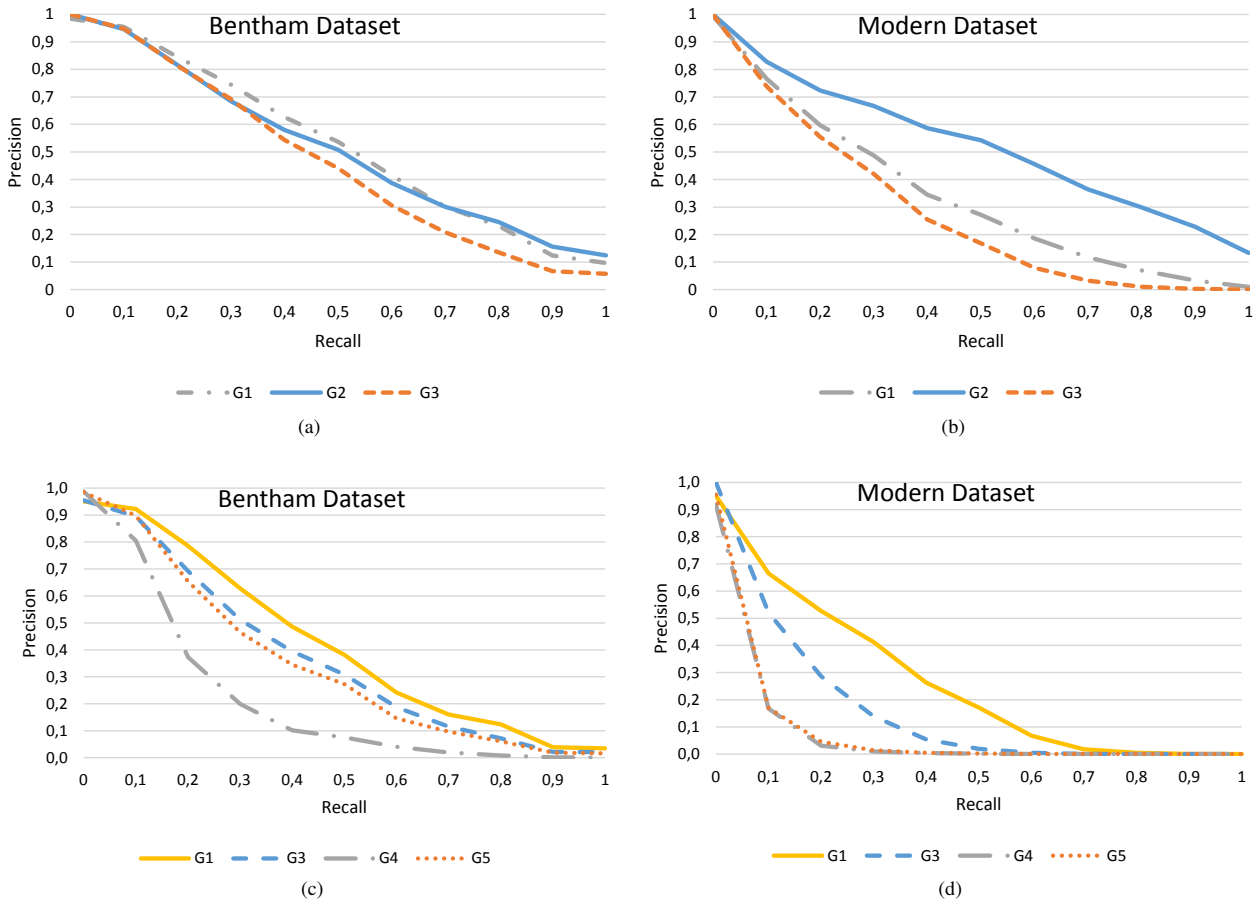


Fig. 6. Precision - Recall Curves for TRACK I: (a) Bentham Dataset, (b) Modern Dataset, TRACK II: (c) Bentham Dataset, (d) Modern Dataset

(FP7/2007-2013) under grant agreement no. 600707 - TranScriptorium. The authors would like to thank the participants for contributing to the realisation of the competition.

REFERENCES

[1] J. Almazán, A. Gordo, A. Fornés, E. Valveny *et al.*, "Handwritten word spotting with corrected attributes," in *ICCV 2013-IEEE International Conference on Computer Vision (2013)*, 2013.

[2] J. Almazán and A. Gordo. (2013) Words with attributes library. [Online]. Available: http://www.cvc.uab.es/~almazan/index.php/projects/words-att/

[3] N. R. Howe, "Part-structured inkball models for one-shot handwritten word spotting," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 582–586.

[4] Y. Leydier, F. Lebourgeois, and H. Emptoz, "Text search for medieval manuscript images," *Pattern Recognition*, vol. 40, no. 12, pp. 3552–3567, 2007.

[5] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *Pattern Recognition*, vol. 42, no. 9, pp. 2089–2105, 2009.

[6] W. Pantke, M. Dennhardt, D. Fecker, V. Märgner, and T. Fingscheidt, "An historical handwritten Arabic dataset for segmentation-free word spotting – HADARA80P," in *Proc. Int. Conf. Frontiers in Handwriting Recognition (ICFHR 2014)*, Crete, Greece, September 2014, accepted.

[7] W. Pantke, V. Märgner, D. Fecker, T. Fingscheidt, A. Asi, O. Biller, J. El-Sana, R. Saabni, and M. Yehia, "HADARA – A software system for semi-automatic processing of historical handwritten Arabic documents," in *Proc. Archiving Conf. 2013*, Washington DC, USA, April 2013, pp. 161–166.

[8] R. Altman, "The illusion of one writer in historical documents and its effect on automating writer identification," in *Proc. Conf. of Int. Graphonomics Society*, Dijon, France, September 2009.

[9] W. Pantke, V. Märgner, and T. Fingscheidt, "On evaluation of segmentation-free word spotting approaches without hard decisions," in *Proc. Int. Conf. Document Analysis and Recognition (ICDAR 2013)*, Washington DC, USA, 2013, pp. 1300–1304.

[10] D. G. Long *et al.*, *The manuscripts of Jeremy Bentham: a chronological index to the collection in the Library of University College, London: based on the catalogue by A. Taylor Milne*. The College, 1981.

[11] B. Gatos, N. Stamatopoulos, and G. Louloudis, "Icdar2009 handwriting segmentation contest," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 14, no. 1, pp. 25–33, 2011.

[12] TREC NIST. (2013) TREC NIST. [Online]. Available: http://trec.nist.gov/pubs/trec16/appendices/measures.pdf

[13] S. A. Chatzichristofis, K. Zagoris, and A. Arampatzis, "The trec files: the (ground) truth is out there," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, ser. SIGIR '11. New York, NY, USA: ACM, 2011, pp. 1289–1290.

[14] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1.

[15] van Rijsbergen C. J., *Information retrieval*, 2nd ed. Butterworths, 1979.