ICDAR 2015 Competition on Keyword Spotting for Handwritten Documents

Joan Puigcerver
PRHLT Research Center
Universitat Politècnica de València
València - Spain
Email: joapuipe@prhlt.upv.es

Alejandro H. Toselli
PRHLT Research Center
Universitat Politècnica de València
València - Spain
Email: ahector@prhlt.upv.es

Enrique Vidal
PRHLT Research Center
Universitat Politècnica de València
València - Spain
Email: evidal@prhlt.upv.es

Abstract—**The principal goal of the Competition on Keyword Spotting for Handwritten Documents, organized under the ICDAR 2015 conference, was to promote different approaches used in the field of Keyword Spotting and to fairly compare them using uniform data and metrics. To accommodate different perspectives adopted by researches in this field, the competition was divided into two distinct tracks, namely, a training-free and a training-based track, and each track entailed two optional assignments. Participants could submit to one or both of them, depending on the capabilities and/or restrictions of their systems. The data used in the competition consisted of historical documents in English with different levels of complexity. This paper presents the details of the competition, including the data, evaluation metrics and results of the best participant methods.**

I. Introduction

Keyword Spotting (KWS) has been considered for both Speech and Text documents under very different points of view and target applications. Perhaps the most recurred distinction is Query-by-Example/Query-by-String (QbE/QbS); i.e., whether the query is by giving a word image example, or just a character string. But many other distinctions are very relevant; among other: Training-based/training-free; i.e., whether the KWS system needs or not to be trained on appropriate (annotated) images, and Segmentation-based/segmentation-free; i.e., whether KWS is applied to full document (page) images or just to images of individual words (previously segmented from the original full images).

Clearly each of these flavors of the KWS problem statement has its own difficulty degree and application targets. In the present contest we aimed at testing, under uniform data sets and benchmark assessment conditions, KWS systems maybe developed under different points of view. This is expected to shear light on the relative capabilities of different approaches and their appropriateness for the different kinds of applications.

The contest was divided into two tracks, and each track consisted of two optional assignments. Participants were able to submit solutions to one or both assignments, depending on the capabilities and/or restrictions of their systems.

In any case, participants had to provide a ranked list, sorted by decreasing confidence, containing the spotted images (segmentation-based assignment) or the precise locations (bounding boxes) where the query words were spotted (segmentation-free assignments).

The taxonomy and characteristics of the different tracks and assignments in the competition are shown below:

1) Track-I: Training-free
    a) Assignment-I.A: Segm-based, QbE
    b) Assignment-I.B: Segm-free, QbE
2) Track-II: Training-based
    a) Assignment-II.A: Segm-free, QbS
    b) Assignment-II.B: Segm-free, QbE

For each assignment, a full baseline system, based on a corresponding well established approach, was provided to the registered groups. The motivation behind these baselines was two-fold:

1) To encourage the participation of researchers in multiple tasks, even if they did not have much expertise from that perspective.
2) To promote new KWS approaches that go significantly beyond the capabilities of traditional and well established methods.

II. Dataset

The dataset consists of a series of handwritten documents written by English philosopher Jeremy Bentham (1748–1832) and its secretaries, prepared by the tranScriptorium project[1]. Some of the documents have already been used in previous KWS and HTR competitions [1], [2], while others have been released for the first time.

The evaluation set consisted of 70 document images, containing several difficult problems to be addressed, including writing from different authors, styles, font-sizes, crossed-out words, etc. We extracted the main block of text from the original pages, so participants did not have to deal with this.

The manually annotated word-segmentation was available only for some pages (57%). For the rest (43%) of pages, a semi-supervised approach was conducted to generate the required bounding boxes: the line-level manual transcripts were force-aligned with the line images, and the resulting bounding boxes were manually corrected for the relevant words. As result, 15 419 segmented word images were obtained from the original 70 pages. See figure 1 for some examples of the document and query images.

[1] http://www.transcriptorium.eu

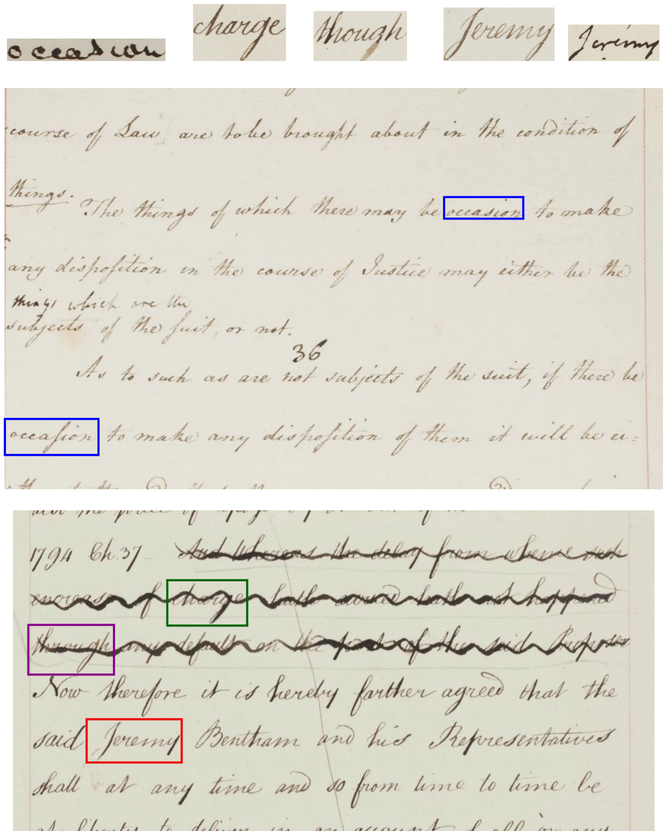

occasion
charge
through
Jeremy
Jeremy
course of Law are to be brought about in the condition of
things. The things of which there may be occasion to make
any disposition in the course of Justice may either be the
things which are the
subjects of the suit, or not.
36
As to such as are not subjects of the suit, if there be
occasion to make any disposition of them it will be e-
1794 Ch. 37
Now therefore it is hereby farther agreed that the
said Jeremy Bentham and his Representatives
shall at any time and so from time to time be

Fig. 1: Examples of the query keywords "OCCASION", "CHARGE", "THROUGH" and "JEREMY", and fragments of two document images containing instances of these words: "OCCASION" in blue, "CHARGE" in green, "THROUGH" in magenta and "JEREMY" in red . Figure better seen in color.

A different set of 423 document images, manually segmented and transcribed into 11 144 lines, was also handed to the participants competing in Track II (training-based track). Only the provided training data was allowed to the participants. We also used this set of pages to extract the query images.

The query set consists of 243 different keyword strings of different lengths (6–15 characters). Each of these strings is represented by 6 or less different example images, making a total of 1421 query images. We ensured that all the query keywords were written at least 4 times in the evaluation set. A particular difficulty added in the selection of the query set is that, all casing instances of a word were considered equivalent, but not plurals or derived words. For instance, "therefore" and "Therefore" are considered the same keyword, but not "according" and "accordingly", nor "instance" and "instances". Figure 2 shows more detailed statistics about the query set.

In addition to the evaluation and the training sets, we also provided a validation set to allow the participants to experiment with the baseline systems and/or to adjust the parameters of their methods. The validation set was significantly smaller than the evaluation set, and contained 10 document images, containing 3 234 words (given also as segmented images for Assignment-I.A). The query set for the validation partition included 95 images of 20 different keywords, extracted from the training page images as well.

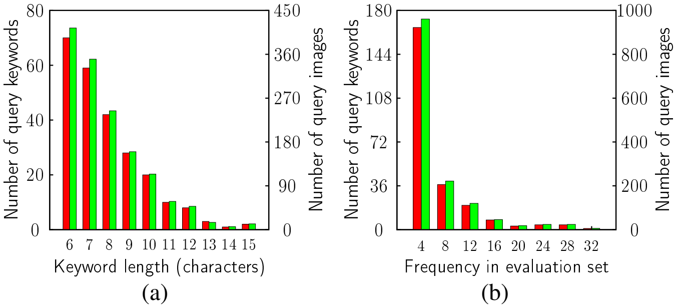


Fig. 2: Histograms for (a) number of query words and images respect the keyword length, and (b) respect their frequency in the evaluation set (bins in the latter histogram are of size four).

III. Evaluation metrics

Mean average precision (mAP) was used to evaluate the solution of each participant. For each query in the set Q, its (interpolated) average precision is computed, using (interpolated) precision-at-top-k, $\pi(k)$ ($\hat{\pi}(k)$), and the recall-at-top-k, $\rho(k)$. Eq. 1 defines the (interpolated) precision and recall scores of the top-k results, using the set of of all relevant items R, and the set of top-k results in the solution $S(k)$; describes the interpolated average precision, AP, where $\Delta\rho(k)$ is the difference in recall between items k and $k-1$; and defines the mAP metric, from the AP of each query q, AP(q).

$$\pi(k) = \frac{R \cap S(k)}{S(k)}; \quad \rho(k) = \frac{R \cap S(k)}{R}; \quad \hat{\pi}(k) = \max_{j:\rho(j)\geq\rho(k)} \pi(j)$$

$$\text{AP} = \sum_{k=1}^{n} \hat{\pi}(k) \cdot \Delta\rho(k); \qquad \text{mAP} = \frac{\sum_{q\in Q} \text{AP}(q)}{|Q|} \tag{1}$$

In segmentation-free scenarios, a result bounding box may not match exactly with the references. Thus, we consider it a correct match when the relative overlapping area with a reference bounding box surpasses a certain threshold (0.7 in this competition), and has the same label as the reference. The overlapping area is computed as follows:

$$O = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

The software implementing the evaluation metrics[2] was given to the participants before the competition, so they could know exactly how their solutions would be evaluated.

Finally, the mAP scores of each assignment were combined to produce a single ranking of participants for each track. First, a score $\text{S}_\text{A}(p,t,a)$ was computed for each participant p, track t, and assignment a, according to Eq. 3. Participants not surpassing the baseline mAP in each assignment, $\text{mAP}_0(t,a)$, were assigned a null score. This thresholding was applied since full baseline implementations (software, scripts, data) were made publicly available from the beginning.

$$\text{S}_\text{A}(p,t,a) = \begin{cases} \frac{\text{mAP}(p,t,a)}{\max_{p'} \text{mAP}(p',t,a)} & \text{mAP}(p,t,a) > \text{mAP}_0(t,a) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

[2] http://www.transcriptorium.eu/~icdar15kws/software/evaluation.py

The scores of both assignments A and B were combined to obtain the score of the participant in each track, according to Eq. 4. The combination rule was designed in order to reward participants with a flexible system without hampering those with a highly-specialized method. The ranking software[3] was also made public beforehand.

$$\begin{aligned} \mathrm{S_T}(p,t) = \quad & \max\{\mathrm{S_A}(p,t,A), \mathrm{S_A}(p,t,B)\} + \\ & 0.2 \cdot \min\{\mathrm{S_A}(p,t,A), \mathrm{S_A}(p,t,B)\} \end{aligned} \tag{4}$$

IV. Baseline methods

Track-I,*[4] *Assignment-I.A. The baseline approach employs MPEG-like descriptors, known as Compact Shape Portrayal Descriptors (CSPD) [3]. Such descriptors are computed from three different features: word image aspect ratio, *Normalized Smoothed Vertical & Horizontal Projections* and, *Normalized Smoothed Top & Bottom Shape Projections*. The *Discrete Cosine Transform* (DCT) is applied on each projection feature by computing only the first 11 coefficients. The first coefficient is actually not considered and used instead for normalizing the values of the ten remaining coefficients. At the end, a final 41-dimensional descriptor is obtained by assembling the coefficients of each of the features, plus the aspect ratio value. Finally, the values of the descriptors are quantized in three bits for binary representation.

In order to measure similarity between images, the weighed Minkowski L1 distance is used, and results are sorted by increasing distance.

Track-I, Assignment-I.B. The baseline method used in this assignment is mainly based on the work described in [4]. The only significant change is that we binarize both query and document images using [5]. Then, both types of images are deskewed, and warping correction and border detection is applied to the document images. Finally, both images are normalized based on the estimated average character height.

Query images are rotated and scaled in order to capture different variations of it. Then, for each generated image, 5 different sets of feature vectors, based on the pixel intensity of non-overlapping windows, are computed. In order to constrain the matching procedure, horizontal RLSA is applied to perform a rough text line estimation. Then a feature vector for each document image is computed, also based on the pixel intensity, on the sections detected by the RLSA algorithm.

Finally, all query feature vectors are matched with the features of the document image, and results are combined in order to produce the final spotting result.

Track-II, Assignments A and B. The baseline systems for both assignments of this track are based on the popular KWS HMM-Filler model [6], originally presented for line-based query-by-string KWS. First, the training line images are preprocessed and a sequence of features is extracted, using a procedure similar to [7]. Then, character HMMs are trained using these features and the corresponding line image transcripts. To speed-up the overall KWS process, we use the HMM-Filler approximation based on character-lattices (CL) [8]: page images are automatically segmented into lines and the CL of each of these lines is obtained and used to compute the likelihood ratio between each keyword-specific HMM and the Filler model. Then, in Assignment-II.A, the standard CL-Filler approach is followed to search for each query string, by using the log likelihood ratio as the confidence that the given keyword is written or not in each test line image.

Since queries in Assignment-II.B are presented in form of images, the keywords in the images are recognized using a character bi-gram language model, trained from the keywords present in the training transcriptions. In order to recognize these images, the processing and feature extraction applied during training is basically applied here. Then, the recognized keyword is simply searched in the CLs as in Assignment-II.A.

In both cases, the bounding boxes are obtained from the line segmentation information and the implicit segmentation given by the HMMs and stored in the CLs.

V. Participant methods

Nine research groups registered to the competition, from which six of them finally participated submitting at least one solution to the evaluation system. Four of them participated in Track-I[5] and the other two in Track-II. Due to limits in the length of this paper, we describe only those systems which overcame the corresponding baselines.

***Pattern Recognition Group (PRG), TU Dortmund University, Germany* – Track I, Assignments A & B** *(Leonard Rothacker, Sebastian Sudholt, Gernot A. Fink)*: Virtually the same approach is used for both assignments. The only difference is that for Assignment-I.B (segmentation-free), an Otsu binarization is performed on the entire page, from which connected components, using binary dilatation, are computed and all segmented, discarding outliers, into word images. From there, the following steps are the same than for Assignment-I.A: Local SIFT descriptors with 4x4 cells spanning 32 pixels, each cell containing 8 bins, are calculated over a dense grid with a step size of 5. A total of $3 \cdot 10^6$ are extracted from all test words and used to calculate a codebook of size 4096. Each word's descriptors are quantized according to this codebook and a two level spatial pyramid is extracted from the quantization. The first level splits the image into 3 equally sized sections along the writing direction while the second level splits up the image in 9 equally sized sections along the same direction.

The query word's descriptors are quantized according to the codebook generated from the test words and spatial pyramids of the layout presented before are extracted. For each query, the test words are ranked based on their Bray-Curtis distance to the corresponding spatial pyramids.

***Computer Vision Center (CVC), Universitat Autònoma de Barcelona, Spain* – Track I, Assignments A & B** *(Marçal Rusiñol, David Aldavert, Alicia Fornés)*: Word images are represented by a descriptor obtained using the Bag-of-Visual-Words (BoVW) framework. A visual representation very similar to the one proposed in [9] was used. First, local regions over the image are densely sampled at a constant step of 3 pixels and at four different scales of sizes of 16, 24, 32 and 40 pixels.

[3] http://www.transcriptorium.eu/~icdar15kws/software/team_ranker.py

[4] We are very thankful to Konstantinos Zagoris and Basilis Gatos for providing the software for baselines I.A and I.B, respectively.

[5] One participant in Track-I asked to withdraw its system from the report.

For the subsequent steps each scale is processed independently. The local regions are characterized by the Integral Histogram of Gradients (IHOG) local descriptor [10] using a $8 \times 4 \times 4$ configuration (128 dimensions). Descriptors are converted into visual words by k-means vector quantization, with a codebook size of 2048. Local descriptors are then encoded into visual words using the Locality-constrained Linear Coding (LLC) algorithm [11] considering the three nearest neighbors.

Spatial information is then added by means of the SPM method [12]. A first level with 3 partitions in the x-axis, and a second level with triple number of divisions in the x-axis and 2 divisions in the y-axis, are used. The visual descriptor is obtained concatenating the histograms of visual words of each pyramid level. This configuration results in a 43008-dimensional visual descriptor per each scale. Then the power normalization proposed in [13] is used in order to increase the response of the visual words with low contribution resulting in a less sparse representation. An L2-normalization is applied to the each scale descriptors, which are subsequently concatenated forming a final 172032-dimensional descriptor which is finally L2-normalized again.

Both word images in the collection and the query images are described as above, and the Euclidean distance is used to obtain a ranked list of results. For each query, the 1400 top-results are returned.

***Computational Intelligence Technology Lab. (CITlab), University of Rostock, Germany* – Track II, Assignments A & B** *(Gundram Leifert, Tobias Strauß, Tobias Grüning, Roger Labahn)*: CITlab's system mainly relies on a recurrent neural network (RNN) based recognition engine, named ARGUS, developed jointly with PLANET intelligent systems GmbH. The overall scheme is essentially the same used previously by CITlab in similar competitions. Please, refer to [14] for additional details. For both assignments in Track-II, the underlying approach is basically the same. The only difference is that for Assignment-II.B (Query-by-Example), a string representation of the query image is first extracted and the Query-by-String approach is followed.

First, text lines are extracted from the provided document images. A rough estimation of the text lines is carried using an algorithm based on the Adaptative RLSA [15]. Then, the final line bounds are extracted using a seam carving approach [16], working on the original input image. Text line images are processed using CITlab's proprietary writing normalization, including contrast and size normalization, and corrections of line bending, slope and slant. This way, all images are re-sized to have 96 pixels height with the writing's main body part appropriately placed and stretched to cover the essential central part of the line image.

The resulting text line images are fed into the RNN with no further segmentation. The output of the RNN estimates the posterior distribution of the alphabet characters, given the whole input image and the position under consideration. The alphabet contains all digits, lowercase and uppercase letters of the standard Latin alphabet, punctuation and other special characters, white-space and a special non-character symbol to detect character boundaries in the output of the RNN.

The output of the RNN is fed into a decoder, which first searches for up to four regions of the line with are likely to include the query keyword. The word may be surrounded by any punctuation mark, white-space or being at the start or end of the line. Details about the specific technology will be subject of upcoming CITlab publications. The candidate parts are scored according and accepted only if its score exceeds a certain threshold, in order to avoid false-positives. Finally, bounding boxes for the words are approximated from the indexes related to the word region and improved using image processing methods.

VI. Results

Table I shows the detailed results of each participant in each one of the two tracks. For the sake of clarity, we have rounded all results to 4 decimal units. We also include the baseline systems in the ranking to show the threshold that the participants had to surpass.

TABLE I: Detailed results and ranking of each participant in (a) Track-I and (b) Track-II. Columns 2 and 4 show the mean average precision, mAP, for each assignment and columns 3 and 5 the *precision-at-5*, $\pi(5)$. Last column shows the final score of the participant in the given track.

(a) Track I: Training-free

Assignment	I.A: Segm. based		I.B: Segm. free		Track-I
Team	mAP	$\pi(5)$	mAP	$\pi(5)$	Score
PRG	**0.4244**	**0.4605**	**0.2761**	**0.3434**	**1.2**
CVC	0.3000	0.3427	0.0821	0.1087	0.7
Baseline	0.1935	0.2241	0.1023	0.1504	—
Withdrawn	0.0024	0.0028	0.0848	0.1088	0.0
CIL	0.1124	0.1475	—	—	0.0

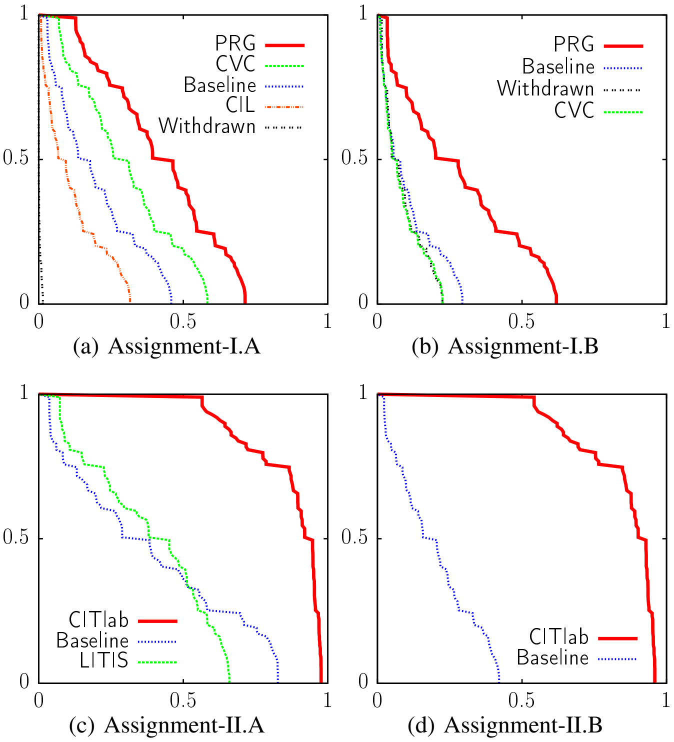
(b) Track II: Training-based

Assignment	II.A: QbS		II.B: QbE		Track-II
Team	mAP	$\pi(5)$	mAP	$\pi(5)$	Score
CITlab	**0.8711**	**0.8737**	**0.8521**	**0.8552**	**1.2**
Baseline	0.3834	0.4831	0.1958	0.2356	—
LITIS	0.3822	0.4864	—	—	0.0

Figure 3 depicts the mean Recall-Precision curves (mRP) achieved by the participants in each assignment. The mRP curves are obtained by averaging the Recall-Precision curves (RP) of each individual query. The area under the mRP curve is the mAP, as the area under the RP curve is the AP. These curves show more details about each solution, such as the maximum precision and recall achieved by each participant.

VII. Conclusions

As it was more or less expected, the competition results confirm that *training-based* methods can achieve much better performance than *training-free* approaches which only rely on knowledge about geometric properties of handwriting images. The Track-II winner method dramatically overcame the performance achieved by all the Track-I (*training-free*) systems, more than doubling the best Track-I mAP under identical test conditions. Their systems excelled in very complicated situations. For instance, they provided accurate bounding boxes even in heavily crossed-out document images regions such as those of the examples shown in Fig. 1.



(a) Assignment-I.A

(b) Assignment-I.B

(c) Assignment-II.A

(d) Assignment-II.B

Fig. 3: Mean Recall-Precision curves (in horizontal and vertical axis, resp.) achieved by the participants and the baseline systems, sorted in decreasing order of mAP. The thickest curve corresponds to the team with the highest mAP in each assignment. Better seen in color.

The CITlab systems were also very much better than the Track-II baselines – note, however, that in order to encourage participation, we used only very simple HMM-based *training-based* baselines (results using more sophisticated HMM-based methods are much closer to those of the winner system).

According to these results, it becomes very evident that, if training data is available, *training-based* are the methods of choice to build systems which achieve practically useful performance. One question remains, however, as to how much training data is actually needed to achieve the bold KWS performance of these methods. Therefore, future competitions in this field should focus on this important aspect to finally help understanding the relative capabilities and requirements of the different approaches to KWS.

Acknowledgment

This work was partially supported by the Spanish MEC under FPU grant FPU13/06281, and under the Valorization and I+D+i Resources program of VLC/CAMPUS, as part of the International Excellence Campus program. It was partially developed also under the STraDA research project (TIN2012-37475-C02-01), by the Generalitat Valenciana under the grant Prometeo/2009/014, and through the EU 7th Framework Programme grant tranScriptorium (Ref:600707). In the organization of this contest we had many fruitful discussions with and received important guidelines from the team which organized the ICFHR-2014 KWS competition. We are grateful to the members of this team for their help. We would finally like to thank all the participants for making this competition a reality.

References

[1] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, and N. Stamatopoulos, “ICFHR 2014 Competition on Handwritten Keyword Spotting (H-KWS 2014),” in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, Sept 2014, pp. 814–819.

[2] J. Andreu Sanchez, V. Romero, A. Toselli, and E. Vidal, “ICFHR 2014 Competition on Handwritten Text Recognition on Transcriptorium Datasets (HTRtS),” in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, Sept 2014, pp. 785–790.

[3] K. Zagoris, K. Ergina, and N. Papamarkos, “Image retrieval systems based on compact shape descriptor and relevance feedback information,” *Journal of Visual Communication and Image Representation*, vol. 22, no. 5, pp. 378 – 390, 2011.

[4] B. Gatos and I. Pratikakis, “Segmentation-free word spotting in historical printed documents,” in *Document Analysis and Recognition, 2009. ICDAR ’09. 10th International Conference on*, July 2009, pp. 271–275.

[5] M. Villegas, V. Romero, and J. A. Sánchez, “On the Modification of Binarization Algorithms to Retain Grayscale Information for Handwritten Text Recognition,” in *7th Iberian Conf. on Pattern Recog. and Image Analysis*, ser. LNCS. Springer, Jun. 2015.

[6] A. Fischer, A. Keller, V. Frinken, and H. Bunke, “Lexicon-free handwritten word spotting using character HMMs,” *Pattern Recognition Letters*, vol. 33, no. 7, pp. 934 – 942, 2012, special Issue on Awards from {ICPR} 2010.

[7] U.-V. Marti and H. Bunke, “Hidden markov models.” River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2002, ch. Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition Systems, pp. 65–90.

[8] A. Toselli and E. Vidal, “Fast HMM-Filler Approach for Key Word Spotting in Handwritten Documents,” in *Document Analysis and Recognition (ICDAR), 2013 12th Int. Conf. on*, Aug 2013, pp. 501–505.

[9] D. Aldavert, M. Rusinol, R. Toledo, and J. Llados, “Integrating Visual and Textual Cues for Query-by-String Word Spotting,” in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, Aug 2013, pp. 511–515.

[10] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, “Fast Human Detection Using a Cascade of Histograms of Oriented Gradients,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 1491–1498.

[11] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2169–2178.

[12] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained Linear Coding for image classification,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 3360–3367.

[13] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” in *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ser. ECCV’10. Springer-Verlag, 2010, pp. 143–156.

[14] G. Leifert, T. Grüning, T. Strauß, G. Leifert, and R. Labahn, “CITlab ARGUS for Keyword Spotting: Description of CITlab’s System for the KWS-2015 Task: Keyword Spotting for Handwritten Documents,” (To be published).

[15] 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, no. 4, pp. 590 – 604, 2010.

[16] N. Arvanitopoulos and S. Susstrunk, “Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts,” in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, Sept 2014, pp. 726–731.