General Overview of ImageCLEF at the CLEF 2016 Labs

Mauricio Villegas[1](✉), Henning Müller[2], Alba G. Seco de Herrera[3], Roger Schaer[2], Stefano Bromuri[4], Andrew Gilbert[5], Luca Piras[6], Josiah Wang[7], Fei Yan[5], Arnau Ramisa[8], Emmanuel Dellandrea[9], Robert Gaizauskas[7], Krystian Mikolajczyk[10], Joan Puigcerver[1], Alejandro H. Toselli[1], Joan-Andreu Sánchez[1], and Enrique Vidal[1]

[1] Universitat Politècnica de València, Spain
mauvilsa@prhlt.upv.es
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] National Library of Medicine, USA
[4] Open University of the Netherlands
[5] University of Surrey, UK
[6] University of Cagliari, Italy
[7] University of Sheffield, UK
[8] Institut de Robòtica i Informàtica Industrial (UPC-CSIC), Barcelona, Spain
[9] École Centrale de Lyon, France
[10] Imperial College London, UK

Abstract. This paper presents an overview of the ImageCLEF 2016 evaluation campaign, an event that was organized as part of the CLEF (Conference and Labs of the Evaluation Forum) labs 2016. ImageCLEF is an ongoing initiative that promotes the evaluation of technologies for annotation, indexing and retrieval for providing information access to collections of images in various usage scenarios and domains. In 2016, the 14th edition of ImageCLEF, three main tasks were proposed: 1) identification, multi-label classification and separation of compound figures from biomedical literature; 2) automatic annotation of general web images; and 3) retrieval from collections of scanned handwritten documents. The handwritten retrieval task was the only completely novel task this year, although the other two tasks introduced several modifications to keep the proposed tasks challenging.

1 Introduction

With the ongoing proliferation of increasingly cheaper devices to capture, amongst others, visual information by means of digital cameras, developing technologies for the storage of this ever growing body of information and providing means to access these huge databases has been and will be an important requirement. As part of this development, it is important to organise campaigns for evaluating the emerging problems and for comparing the proposed techniques for solving the problems based on the exact same scenario in a reproducible way.

Motivated by this, ImageCLEF has for many years been an ongoing initiative that aims at evaluating multilingual or language independent annotation and retrieval of images [20]. The main goal of ImageCLEF is to support the advancement of the field of visual media analysis, classification, annotation, indexing and retrieval, by proposing novel challenges and developing the necessary infrastructure for the evaluation of visual systems operating in different contexts and providing reusable resources for benchmarking. Many research groups have participated over the years in its evaluation campaigns and even more have acquired its datasets for experimentation. The impact of ImageCLEF can also be seen by its significant scholarly impact indicated by the substantial numbers of its publications and their received citations [30].

There are other evaluation initiatives that have had a close relation with ImageCLEF. LifeCLEF [16] was formerly an ImageCLEF task. However, due to the need to assess technologies for automated identification and understanding of living organisms using data not only restricted to images, but also videos and sound, it was decided to be organised independently from ImageCLEF. Other CLEF labs linked to ImageCLEF, in particular the medical task, are: CLEFeHealth [11] that deals with processing methods and resources to enrich difficult-to-understand eHealth text and the BioASQ [2] tasks from the Question Answering lab that targets biomedical semantic indexing and question answering. Due to their medical topic, the organisation is coordinated in close collaboration with ImageCLEF. In fact, at CLEF 2015 there was a joint session on the "Challenges and synergies in the evaluation of health IR/IE".

This paper presents a general overview of the ImageCLEF 2016 evaluation campaign[1], which as usual was an event organised as part of the CLEF labs[2]. Section 2 presents a general description of the 2016 edition of ImageCLEF, commenting about the overall organisation and participation in the lab. Followed by this are sections dedicated to the three main tasks that were organised this year, Section 3 for the medical task that deals mainly with compound figures from biomedical literature and how to make their visual content accessible, Section 4 for the image annotation task, and Section 5 for the new task introduced this year targeted at the retrieval from scanned handwritten document collections. These sections are only short summaries of the tasks. For the full details and complete results, the readers should refer to the the corresponding task overview papers [15,9,36]. The final section of this paper concludes by giving an overall discussion, and pointing towards the challenges ahead and possible new directions for future research.

2 Overview of Tasks and Participation

The 2016 edition of ImageCLEF consisted of three main tasks that covered challenges in diverse fields and usage scenarios. In 2015 [31] all the tasks addressed

[1] http://imageclef.org/2016/

[2] http://clef2016.clef-initiative.eu/

topics related to processing the images in order to automatically assign metadata to them, not directly evaluating retrieval, but techniques that produce valuable annotations that can be used for subsequent image database indexing, mining or analysis. This year there was also a new task that evaluated retrieval of small segments from images containing handwritten text. The three tasks organised were the following:

- **Medical Classification:** addresses the identification, multi-label classification, caption prediction and separation of compound figures commonly found in the biomedical literature.
- **Image Annotation:** aims at developing systems for automatic annotation of concepts, their localization within the image, and generation of sentences describing the image content in a natural language. A pilot task on text illustration is also introduced in this edition.
- **Handwritten Scanned Document Retrieval:** targets the challenge of retrieval of page segments in scanned handwritten documents for multi-word textual queries.

The medical and annotation tasks were continuations from previous years, however, both introduced changes. In comparison to 2015, the medical task provided a larger amount of data with more training data and also introduced a new subtask of which the objective was the prediction of figure captions given the image, so providing 5 subtasks. The photo annotation task was also changed, having 4 subtasks this year, two of which were continued from last year and two new ones: selection of concepts for inclusion in generated image descriptions, and finding the best image to illustrate a given text snippet.

In order to participate in the evaluation campaign, the groups first had to register either on the CLEF website or from the ImageCLEF website. To actually get access to the datasets, the participants were required to submit a signed End User Agreement (EUA) by email. Table 1 presents a table that summarize the participation in ImageCLEF 2016, including the number of registrations and number of signed EUAs, indicated both per task and for the overall lab. The table also shows the number of groups that submitted results (a.k.a. runs) and the ones that submitted a working notes paper describing the techniques used.

The number of registrations could be interpreted as the initial interest that the community has for the evaluation. However, it is a bit misleading because several people from the same institution might register, even though in the end they count as a single group participation. The EUA explicitly requires all groups that get access to the data to participate. Unfortunately, the percentage of groups that submit results is often relatively small. Nevertheless, as observed in studies of scholarly impact [30], in subsequent years the datasets and challenges provided by ImageCLEF do get used quite often, which in part is due to the researchers that for some reason were unable to participate in the original event.

Although for the 2015 edition of ImageCLEF the participation increased considerably with respect to previous years, this was no longer the case for the current 2016 edition. This could be in part due to a CLEF restriction that required to reduce the number of tasks from four to three. However, the number

Table 1: Key figures of participation in ImageCLEF 2016.

Task	Online registrations	Signed EUA	Groups that subm. results	Submitted working notes
Medical Classification	46	24	8	5
Image Annotation	53	28	7	7
Handwritten Retrieval	48	24	4	3
Overall[⋆]	98	54	19	15

[⋆] Unique groups. None of the groups participated in multiple tasks.

of registrations and signed EUAs for the continuing tasks also decreased. The new handwritten retrieval task had quite a large number or registrations and EUAs, comparable to the other tasks. In fact, 13 groups signed the EUA only for this task, giving the impression that there is a considerable interest in this area.

The following three sections are dedicated to each of the tasks. Only a short overview is reported, including general objectives, description of the tasks and datasets and a short summary of the results.

3 The Medical Task

An estimated over 40% of the figures in the medical literature in PubMed Central are compound figures (images consisting of several subfigures) [13] like the images in Figure 1. The compound figures in the articles are made available in a single block and are not separated into subfigures. The figures contain diverse information and often subfigures of various image types or modalities. Therefore, being able to separate and/or label each of the figures can help image retrieval systems to focus search and deliver focused results. For more details on this task please refer to the task overview paper [15].

3.1 Past Editions

Since 2004, ImageCLEF has run the medical task, ImageCLEFmed, to promote research on biomedical image classification and retrieval [17]. ImageCLEFmed has evolved strongly to adapt to the current challenges in the domain. The objective of ImageCLEFmed 2015 [14] and 2016 has been to work in large part on compound figures of the biomedical literature and to separate them if possible and/or attach to the subparts labels about the content. In 2013 a compound figure separation subtask was already introduced as a pilot task. A totally new subtask to predict image captions was introduced in 2016. The objective is also to create manually labelled resources on the many images in PubMed Central.

3.2 Objectives and Subtasks for the 2016 Edition

The novelties introduced in the tasks for 2016 are the distribution of a larger number of compound figures compared to the previous years and the introduction of the caption prediction subtask. Thus, there were five types of subtasks in 2016:

- **Compound figure detection**: This subtask was first introduced in 2015. Compound figure identification is a required first step to separate compound figures from images with a single content. Therefore, the goal of this subtask is to identify whether a figure is a compound figure or not. The subtask makes training data available containing compound and non compound figures from the biomedical literature.
- **Multi-label classification**: This subtask was first introduced in 2015. Characterization of the content in compound figures is difficult, as they may contain subfigures from various imaging modalities or image types. This subtask aims to label each compound figure with each of the image types (of the 30 classes of a defined hierarchy [21]) of the subfigures contained without knowing where the separation lines are.
- **Figure separation**: This subtask was first introduced in 2013. The subtask makes available training data with separation labels of the figures and then a test data set where the labels are made available after the submission of the results for the evaluation. Evaluation is not based on strict placement of separation lines but on proximity to separation lines.
- **Subfigure classification**: This subtask was first introduced in 2015 but similar to the modality classification subtask organized in 2011-2013. This

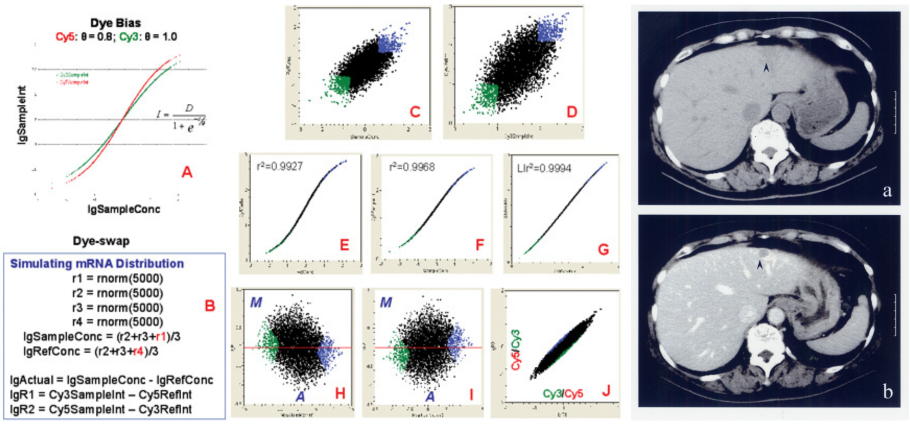


Fig. 1: Examples of compound figures in the biomedical literature.

subtask aims to classify images into the 30 classes of the image type hierarchy. The images are the subfigures extracted from the compound figures distributed for the multi-label subtask.

- **Caption prediction**: This is a new subtask that was introduced in 2016. The subtask aims to evaluate algorithms that can predict captions for the diagnostic images provided as training and test set. The performance is measured based on word similarity between predictions and real captions.

3.3 Participation and Results

Table 1 shows the participation in this task. In 2016, there were slightly fewer registrations than in 2015, however the same number of groups submitted runs and the total number of submitted runs increased to 69.

Three groups participated in the compound figure detection subtask. The DUTIR group obtained the best results achieving a 92.7% of accuracy (see Table 2). Multi-label classification had two participants BMET and MLKD. BMET has achieved the best combined result of an Hamming loss of 1.35% and an f-measure of 32% (see Table 3). Only one participant, UDEL CIS [26], submitted 10 runs to the figure separation subtask with an accuracy of 84.43% (see Table 4). The subfigure classification subtask was the most popular with 42 runs

Table 2: Results of the best runs of the compound figure detection task.

Group	Run	Run Type	Accuracy
DUTIR	CFD_DUTIR_Mixed_AVG	mixed	92.70
CIS UDEL	CFDRun10	mixed	90.74
MLKD	CFD2	textual	88.13
DUTIR	CFD_DUTIR_Visual_CNNs	visual	92.01

Table 3: Best runs of the figure separation task.

Group	Run	Hamming Loss	F-Measure
BMET	1462019299651__MLC-BMET-multiclass-test-max-all	0.0131	0.295
BMET	1462019365690__MLC-BMET-multiclass-test-prob-max-all	0.0135	0.32
MLKD	1462024417416__MLC2	0.0294	0.32

Table 4: Best runs of the figure separation task.

Run	Group	Run Type	Accuracy
CIS UDEL	FS.run9	visual	84.43

submitted. BCSG [18] achieved the best results with an accuracy of 88.43%, a good increase compared to past years. Unfortunately, there were no participants in the caption prediction task, however the data are made available and will hopefully be used in the future. A more detailed analysis of the medical classification tasks including tables with results of all runs is presented in the task overview paper of the working notes [15].

Table 5: Results of the best runs of the subfigure classification task.

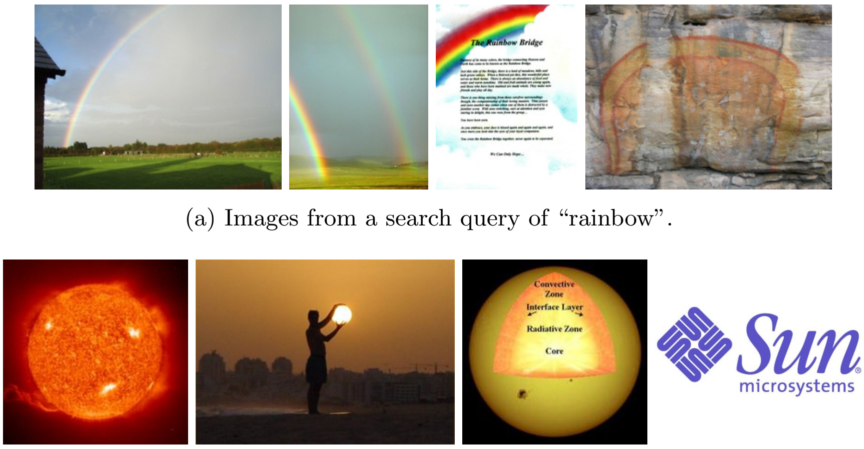
Run	Group	Run Type	Accuracy
BCSG	SC_BCSG_run10_Ensemble_Vote	mixed	88.43
BCSG	SC_BCSG_run2_Textual	textual	72.22
MLKD	SC2	textual	58.37
BCSG	SC_BCSG_run8_DeCAF_ResNet-152_PseudoInverse	visual	85.38
BCSG	SC_BCSG_run1_Visual	visual	84.46
IPL	SC_enriched_GBOC_1x1_256_RGB_Phow_Default_1500_EarlyFusion	visual	84.01
BMET	SC-BMET-subfig-test-prob-sum	visual	77.55
CIS UDEL	SCRun1	visual	72.46
NWPU	sc.run2	visual	71.41
NOVASearch	SC_NOVASearch_cnn_10_dropout_vgglike.run	visual	65.31

4 The Image Annotation Task

Since 2010, ImageCLEF has run a scalable concept image annotation task to promote research into the annotation of images using large-scale, noisy web page data in a weakly-supervised fashion. The main motivation for the task comes from the large number of mixed-modality data (e.g. web page text and images) which can be gathered cheaply from the Internet. Such data can potentially be exploited for image annotation. Thus, the main goal of the challenge is to encourage creative ideas of using such noisy web page data to improve various image annotation tasks: localizing different concepts depicted in images, generating descriptions of the scenes, and text-to-image retrieval.

4.1 Past Editions

The Scalable Concept Image Annotation task is a continuation of the general image annotation and retrieval task that has been held every year at ImageCLEF since its very first edition in 2003. In the first editions the focus was on retrieving images relevant to given (multilingual) queries from a web collection, while from 2006 onwards annotation tasks were also held, initially aimed at object detection, but more recently also covering semantic concepts. In its current form, the 2016 Scalable Concept Image Annotation task [9] is its fifth edition, having been organized in 2012 [32], 2013 [34], 2014 [33], and 2015 [8]. In the 2015 edition [8], the image annotation task was expanded to concept localization and also natural language sentential description of images. In the 2016 edition, we further



(a) Images from a search query of "rainbow".

(b) Images from a search query of "sun".

Fig. 2: Examples of images retrieved by a commercial image search engine.

introduced a text illustration 'teaser' task[3], to evaluate systems that analyze a text document and select the best illustration for the text from a large collection of images provided. As there is an increased interest in recent years in research combining text and vision, the new tasks introduced in both the 2015 and 2016 editions aim at further stimulating and encouraging multimodal research that use both text and visual data for image annotation and retrieval.

4.2 Objective and Task for the 2016 Edition

Image annotation has generally relied on training data that are manually, and thus reliably annotated. Annotating training data is an expensive and laborious endeavour that cannot easily scale, particularly as the number of concepts grow. However, images for any topic can be cheaply gathered from the Web along with associated text from the web pages that contain the images. The degree of relationship between these web images and the surrounding text varies greatly, i.e., the data are very noisy but overall these data contain useful information that can be exploited to develop annotation systems. Figure 2 shows examples of typical images found by querying search engines. As can be seen, the data obtained are useful and furthermore a wider variety of images is expected, not only photographs, but also drawings and computer generated graphics. Likewise there are other resources available that can help to determine the relationships between text and semantic concepts, such as dictionaries or ontologies.

[3] A second teaser task was actually also introduced, aimed at evaluating systems that identify the GPS coordinates of a text document's topic based on its text and image data. However, we had no participants for this task, and thus will not discuss the second teaser task in this paper.

The goal of this task is to evaluate different strategies to deal with the noisy data so that it can be reliably used for annotating, localizing, generating natural sentences and retrieving images from practically any topic. As in the 2015 task, external training data such as ImageNet ILSVRC2015 and MSCOCO are allowed and encouraged. However, in contrast to previous years, in this edition participants are expected to produce two sets of related results:

1. One approach using only externally trained data;
2. The second approach using both external data and the noisy web data of 510,123 web pages.

The motivation for this is to encourage participants to utilize the provided 510,123 web pages to improve the performance of systems trained using external data. This also distinguishes the ImageCLEF image annotation task from other similar image annotation challenges. This year's challenge comprises four subtasks:

1. **Subtask 1: Image Annotation and Localization.** This subtask required participants to develop a system that receives as input an image and produces as output a prediction of which concepts are present in that image, selected from a predefined list of concepts, and where they are located within the image. Participants were requested to annotate and localize concepts in all 510,123 images.
2. **Subtask 2: Natural Language Caption Generation.** This subtask required the participants to develop a system that receives as input an image and produces as annotation a sentential, textual description of the visual content depicted in the image. Again, the test set is all 510,123 images.
3. **Subtask 3: Content Selection.** This subtask is related to Subtask 2, but is aimed primarily at those interested only in the natural language generation aspects of the task. It concentrates on the content selection phase of image description generation, i.e. which concepts should be selected to be mentioned in the corresponding description? Gold standard input (bounding boxes labelled with concepts) is provided for each of the 450 test images, and participants are expected to develop systems that predict the bounding box instances most likely to be mentioned in the corresponding image descriptions. Unlike the 2015 edition, participants were not required to generate complete sentences, but were only requested to provide a list of bounding box instances per image.
4. **Teaser task: Text Illustration.** The teaser task is designed to evaluate the performance of methods for text-to-image matching. Participants were asked to develop a system to analyse a given text document and find the best illustration for it from a set of all available images. The 510,123 dataset was split into 310,123 and 200,000 documents for training and testing respectively. At test time, participants were provided as input 180,000 text documents extracted from a subset of the test documents as queries, and the goal is to select the best illustration for each text from the 200,000 test images.

Table 6: Results for Subtask 1: Image Annotation and Localization.

Group	**0% Overlap**	**50% Overlap**
CEA	0.54	0.37
MRIM	0.25	0.14
CNRS	0.20	0.11
UAIC	0.003	0.002

The concepts this year (for the main subtasks) were retained from the 2015 edition. They were chosen to be visual objects that are localizable and that are useful for generating textual descriptions of visual content of images, including animated objects (person, dogs), inanimated objects (houses, cars) and scenes (city, mountains).

The noisy dataset used in this task was based on the 2015 edition with 500,000 documents. In the 2016 edition, the dataset was augmented with approximately 10,123 new image-document pairs from a subset of the BreakingNews dataset [24] which we developed, bringing the total number of documents to approximately 510,123. However, the subset of the data used for evaluating the three main subtasks remains the same, thus making the evaluation process comparable to the 2015 edition.

4.3 Participation and Results

In 2016, 7 groups participated in the task, submitting over 50 runs across the subtasks, and all 7 also produced working notes.

Four teams submitted results in Subtask 1 to produce localized predictions of image concepts on images. The subtask was evaluated using the PASCAL VOC style metric of intersection over union (IoU), the area of intersection between the foreground in the output segmentation and the foreground in the ground-truth segmentation, divided by the area of their union. The final results are presented in Table 6 in terms of mean average precision (MAP) over all images of all concepts, with both 0% overlap (i.e. no localization) and 50% overlap. The method of computing the performance was adjusted from the previous year, it now includes recall at a concept level, penalising approaches that only detect a few concepts (for example face parts) by averaging the precision overall concepts. This has reduced the overall scores, however if the approaches are analysed using last years evaluation method, the approach by CEA, has increased by around 8%, indicating progress is continuing in this area. All approaches use a deep learning framework, including a number using the Deep Residual Convolutional Neural Network (ResNet) [12]. This explains and verifies much of the improvement over previous years. Face detection was fused into a number of approaches, however, in general, it was not found to provide much improvement in comparison to the improved neural network. A shortcoming of the challenge, however, is still present and with increasing performance is being a larger problem. There is a difficulty in ensuring that the ground truth has 100% of the concepts labelled, thus allowing a recall measure to be used. The current crowdsourcing-based hand labelling of the ground truth is found to not achieve this and so a recall measure is not evaluated.

Two teams participated in Subtask 2 to generate natural language image descriptions. The subtask was evaluated using the Meteor evaluation metric [4]. Table 7 shows the Meteor scores for the best run for each participant. ICTisia achieved the better Meteor score of 0.1837 by fine-tuning on the state-of-the-art joint CNN-LSTM image captioning system. UAIC who also participated last year improved their score with 0.0934 compared to their performance from last year (0.0813), using a template-based approach to the problem.

Table 7: Results for Subtask 2: Natural Language Caption Generation.

Team	**Meteor**
ICTisia	0.1837 ± 0.0847
UAIC	0.0934 ± 0.0249

Subtask 3 on content selection was also represented by two teams. The subtask was evaluated using the fine-grained metric proposed for last year's challenge [8,37]. Table 8 shows the F-score, Precision and Recall across 450 test images for each participant. DUTh achieved a higher F-score compared to the best performer from last year (0.5459 vs. 0.5310), by training SVM classifiers given various image descriptors. While UAIC did not significantly improve their F-score from last year, their recall score shows improvement.

Table 8: Results for Subtask 3: Content Selection.

Team	**Content Selection Score**		
	Mean F	**Mean** P	**Mean** R
DUTh	0.5459 ± 0.1533	0.4451 ± 0.1695	0.7914 ± 0.1960
UAIC	0.4982 ± 0.1782	0.4597 ± 0.1553	0.5951 ± 0.2592

Table 9 shows the result of the pilot teaser task of text illustration. This task yielded interesting results. Bearing in mind the difficulty of the task (selecting one correct image from 200,000 images), CEA yielded a respectable score that is much better than chance performance, by mapping visual and textual modalities onto a common space and combining this with a semantic signature. INAOE on the other hand produced superior results with a bag-of-words approach. Both teams performed better on the larger 180K test set than the more restricted 10K test set (news domain), although INAOE performed better on the 10K test set at smaller ranks (1-10).

Table 9: Results for Teaser task: Text Illustration. The Recall@K are shown for each participant's best run, for a selected subset of the test set (10K) and the full test set (180K).

Team	**Test set**	**Recall (%)**					
		R@1	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
Random Chance	-	0.00	0.00	0.01	0.01	0.03	0.05
CEA	10K	0.02	0.05	0.11	0.26	0.46	0.80
	180K	0.18	0.63	1.05	1.97	3.00	4.51
INAOE	10K	37.05	73.12	78.06	79.55	79.74	79.77
	180K	28.75	63.50	75.48	84.39	86.79	87.59

For a more detailed analysis and discussion of the results, please refer to the task overview paper [9].

5 The Handwritten Retrieval Task

In recent years there has been an increasing interest in digitising the vast amounts of pre-digital age books and documents that exist throughout the world. Many of the emerging digitisation initiatives are for huge collections of handwritten documents, for which automatic recognition is not yet as mature as for printed text Optical Character Recognition (OCR). Thus, there is a need to develop reliable and scalable indexing techniques for manuscripts, targeting its particular challenges. Users for this technology could be libraries with fragile historical books, which for preservation are being scanned to make them available to the public without the risk of further deterioration. Apart from making the scanned pages available, there is also great interest in providing search facilities so that the people consulting these collections have information access tools that they are already accustomed to. The archaic solution is to manually transcribe and then use standard text retrieval technologies. However, this becomes too expensive for large collections. Alternatively, handwritten text recognition (HTR) techniques can be used for automatic indexing, which requires to transcribe only a small part of the document for training the models, or reuse models obtained from similar manuscripts, thus requiring the least human effort.

5.1 Previous Work

Traditionally the task of searching in handwritten documents has been known as *Keyword Spotting* (KWS), which actually can be seen as a particular case of image retrieval. The goal of KWS is to find all instances of a query in a given document. Among the noteworthy KWS paradigms aiming to fulfil this goal, two main kinds are distinguished: *Query by Example* (QbE) [10,7,1,38] and *Query by String* (QbS) [5,6,25,29]. While in QbE a query is specified by providing a text image to be searched for, in QbS, queries are directly specified as character strings. Likewise other distinctions considered are: training-based/free [5,38]; i.e., whether the KWS system needs or not to be trained on appropriate (annotated) images, and segmentation-based/free [38,7]; 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).

In the last years, several KWS contests on handwritten documents have been organised, mainly within the frame of conferences like ICFHR and ICDAR. These focused first on benchmarking QbE approaches [22][4,5], although lately, QbS approaches have also been considered in the ICDAR'15 [23][6], and in the ICFHR'16[7].

Regarding literature about how to deal with hyphenated words, it is worth mentioning the approaches described in [19,27,28].

5.2 Objective and Task for the 2016 Edition

The task targeted the scenario of free text search in a set of handwritten document images, in which the user wants to find sections of the document for a given multiple word textual query. The result of the search is not pages, but smaller regions (such as a paragraph), which can even include the end of a page and start of the next. The system should also be able to handle words broken between lines and words that were not seen in the data used for training the recognition models. Figure 3 shows an example search result that illustrates the intended scenario.

Since the detection of paragraphs is in itself difficult, to simplify the problem somewhat the segments to retrieve were defined to be a concatenation of 6 consecutive lines (from top to bottom and left to right if there page had columns), ignoring the type of line it may be (e.g. title, inserted word, etc.). More precisely, the segments are defined by a sliding window that moves one line at a time (thus neighbouring segments overlap by 5 lines) traversing all the pages in the document, so there are segments that include lines at the end of a page and at the beginning of the next.

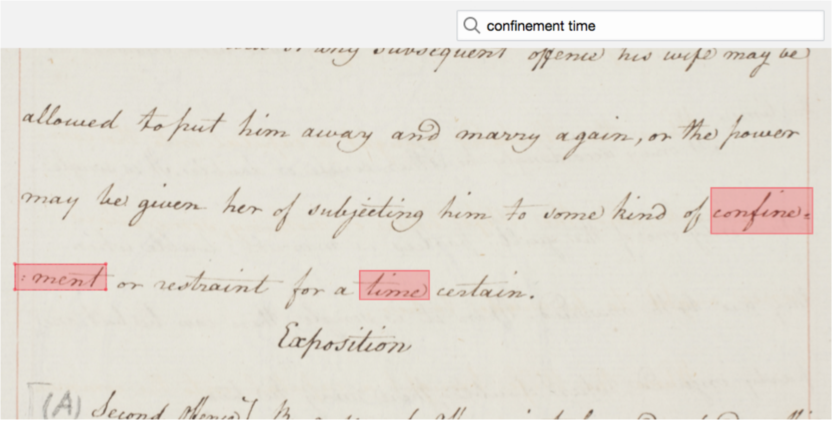
The queries were one or more words that had to be searched for in the collection, and a segment is considered relevant if all the query words appear in the given order. The participants were expected to submit for each query, only for the segments considered relevant, a relevance score and the bounding boxes of all appearances of the query words within the segment irrespectively if it was or not an instance of the word that made the segment relevant. The queries were

[4] http://users.iit.demokritos.gr/~bgat/H-WSCO2013

[5] http://vc.ee.duth.gr/H-KWS2014

[6] http://transcriptorium.eu/~icdar15kws

[7] https://www.prhlt.upv.es/contests/icfhr2016-kws



confinement time

any subsequent offence his wife may be
allowed to put him away and marry again, or the power
may be given her of subjecting him to some kind of confine-
-ment or restraint for a time certain.
Exposition
[(A) Second offence]

Fig. 3: Example of a page segment search result in a handwritten document.

selected such that key challenges were included: words broken between lines, queries with words not seen in the training data, queries with repeated words, and queries with zero relevant results.

The dataset used in the task was a subset of pages from unpublished manuscripts written by the philosopher and reformer, Jeremy Bentham, that were digitised and transcribed under the Transcribe Bentham project [3]. All of the provided data for the task and scripts for computing the evaluation measures and the baseline system are publicly available and citable [35].

These kinds of evaluations related to handwriting recognition are normally organised in conjunction with more specialised conferences such as ICDAR and ICFHR. The reason for organising it at CLEF was that most of the research done up to now in this area does not address all challenges from the perspective of information retrieval. So the objective was to disseminate these problems to experts from the information retrieval community so that they get more involved. Thus, the task was designed to allow easy participation from different research communities by providing prepared data for each, with the aim of having synergies between these communities, and providing different ideas and solutions to the problems being addressed. The original page images were provided, so that the all parts of the task could be addressed, including extraction of lines, preprocessing of the images, training of recognition models, decoding, indexing and retrieval. Also recognition results for a baseline system were provided in plain text so that groups working on text retrieval could participate without worrying about images. Finally the training set included bounding boxes automatically obtained for all of the words, so that groups working on query-by-example keyword spotting could participate, although with the twist that the example images could be incorrectly segmented, so a technique to select the among the available example words would be required.

For further details on the task, results and data please refer to the overview paper [36] and/or the dataset repository [35].

Table 10: Summary of results (in %) for the Handwritten Retrieval task.

Group	AP		mAP		NDCG		mNDCG	
	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test
CITlab	95.0	47.1	89.8	39.9	96.8	62.7	90.9	41.7
IIIT	41.5	3.4	22.5	3.4	49.4	8.8	26.1	3.9
MayoBMI	25.8	2.5	23.4	2.9	33.1	7.0	26.6	3.6
UAEMex	61.1	0.3	38.5	0.4	69.1	1.2	41.7	0.4
Baseline	74.2	14.4	49.9	8.1	80.1	27.5	51.7	9.4

5.3 Participation and Results

There was considerable interest in the task. Over twenty groups signed the EUA, and based on the data server logs, the test queries (only useful if there was an intention of submitting results) were downloaded from 9 countries: Germany, India, Israel, Japan, Mexico, Morocco, Tunisia, Turkey and USA. In the end, only four groups submitted results and three of them submitted a working notes paper describing their system.

Table 10 presents for each group that participated, the best result both for the development and test sets, and including only the segment based performance measures, i.e., does not consider the predicted word bounding boxes. The assessment uses the Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG), measured both globally or as the mean (preceded by a lower-case m) for all evaluated queries.

Each group followed quite a different approaches. The IIIT team participated as query by example, thus their results are not directly comparable with any of the others. Two teams, MayoBMI and UAEMex, based their work on the provided recognition results, although considered only the 1-best, thus being limited in comparison to the baseline system. Furthermore, the test set was considerably more difficult than the development and the baseline system performed poorly, so their results were also affected by this. Two groups, CITlab and MayoBMI, dealt with the broken words, though both based it on the detection of hyphenation symbols, even thought there could be broken words without any hyphenation. The MayoBMI did not finally submit results with hyphenation detection since they considered the performance insufficient. Only the CITlab group tackled the complete problem, training recognition models and retrieving broken words and words unseen in training. They also used Recurrent Neural Networks, which is the current state of the art in handwriting recognition, which clearly reflects in the obtained results.

For the complete results, including specific analysis of the words unseen in training and the broken words, the reader is invited to read the task specific overview paper [36].

6 Conclusions

This paper presents a general overview of the activities and outcomes of the 2016 edition of the ImageCLEF evaluation campaign. Three main tasks were organised covering challenges in: identification, multi-label classification, caption prediction and separation of compound figures from biomedical literature; automatic concept annotation, localization, sentence description generation and retrieval of web images; and retrieval of page segments in handwritten documents.

The participation was similar to the 2013 and 2014 editions, although it decreased with respect to the 2015 edition, in which the participation was outstandingly high. Nineteen groups submitted results and fifteen of them provided a working notes paper describing their work. Even though the participation was not as good as hoped, the obtained results are interesting and useful.

Several new challenges in the medical tasks were provided focusing on the challenges dealing with compound figures. Many groups now employed deep learning algorithms or mixed handcrafted approaches with deep learning. Results obtained were very good in several of the tasks showing a general increase in the quality of the algorithms.

The image annotation challenges indicate the mainstream acceptance of deep neural networks, with much of the improvement in subtask 1 being provided by improved neural networks. Several groups used a face detection to improve results, however the text analysis for image annotation has in general been dropped at the moment, due to the neural network improvements. In subtask 2, one team also utilised the state-of-the-art neural network based image captioning system, while the others used a conventional template-based approach. Subtask 3 on the other hand relied on conventional techniques such as SVMs, due to the smaller development set. Interesting, a simple bag-of-words approach yielded significantly better results in the large-scale text illustration teaser task compared to neural network based techniques.

In the new task related to handwritten retrieval, very good results were obtained by one of the participants, in particular handling moderately well the novel challenge of retrieving words broken between lines. The other groups did not obtain optimal results, but tried interesting ideas for working with the automatic recognition of the images in order to index them. The produced dataset and proposed challenges surely will serve as basis for future works and evaluations.

ImageCLEF brought again a together an interesting mix of tasks and approaches and we are looking forward to the discussions at the workshop.

Acknowledgements

The general coordination and the handwritten retrieval task have been supported by the European Union (EU) Horizon 2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943), EU project HIMANIS (JPICH programme, Spanish grant Ref: PCIN-2015-068) and MINECO/FEDER, UE under project TIN2015-70924-C2-1-R. The image annotation task is co-organized by the VisualSense (ViSen) consortium under the ERA-NET CHIST-ERA D2K 2011 Programme, jointly supported by UK EPSRC Grants EP/K01904X/1 and EP/K019082/1, French ANR Grant ANR-12-CHRI-0002-04 and Spanish MINECO Grant PCIN-2013-047.

References

1. Aldavert, D., Rusinol, M., Toledo, R., Llados, J.: Integrating Visual and Textual Cues for Query-by-String Word Spotting. In: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. pp. 511–515 (Aug 2013)
2. Balikas, G., Kosmopoulos, A., Krithara, A., Paliouras, G., Kakadiaris, I.A.: Results of the bioasq tasks of the question answering lab at CLEF 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. (2015)
3. Causer, T., Wallace, V.: Building a volunteer community: results and findings from Transcribe Bentham. Digital Humanities Quarterly 6 (2012), http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html
4. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation (2014)
5. Fischer, A., Keller, A., Frinken, V., Bunke, H.: Lexicon-free handwritten word spotting using character HMMs. Pattern Recognition Letters 33(7), 934 – 942 (2012), special Issue on Awards from {ICPR} 2010
6. Frinken, V., Fischer, A., Bunke, H.: A Novel Word Spotting Algorithm Using Bidirectional Long Short-Term Memory Neural Networks. In: Schwenker, F., El Gayar, N. (eds.) Artificial Neural Networks in Pattern Recognition, Lecture Notes in Computer Science, vol. 5998, pp. 185–196. Springer Berlin / Heidelberg (2010)
7. Gatos, B., Pratikakis, I.: Segmentation-free word spotting in historical printed documents. In: Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on. pp. 271–275 (July 2009)
8. Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Toulouse, France (September 8-11 2015)
9. Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 5-8 2016)
10. Giotis, A., Gerogiannis, D., Nikou, C.: Word Spotting in Handwritten Text Using Contour-Based Models. In: Frontiers in Handwriting Recog. (ICFHR), 2014 14th Int. Conf. on. pp. 399–404 (Sept 2014)
11. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J.R.M., Zuccon, G.: Overview of the CLEF ehealth evaluation lab 2015. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings. pp. 429–443 (2015), doi:10.1007/978-3-319-24027-5_44

12. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: The IEEE International Conference on Computer Vision (ICCV) (December 2015)
13. García Seco de Herrera, A., Kalpathy-Cramer, J., Demner Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. In: Working Notes of CLEF 2013 (Cross Language Evaluation Forum) (September 2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-SecoDeHerreraEt2013b.pdf
14. García Seco de Herrera, A., Müller, H., Bromuri, S.: Overview of the ImageCLEF 2015 medical classification task. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum). CEUR Workshop Proceedings, CEUR-WS.org (September 2015)
15. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 Medical Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 5-8 2016)
16. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W., Planquè, R., Rauber, A., Palazzo, S., Fisher, B., Müller, H.: Lifeclef 2015: Multimedia life species identification challenges. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings. pp. 462–483 (2015), doi:10.1007/978-3-319-24027-5_46
17. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems –an overview of the medical image retrieval task at imageclef 2004-2014. Computerized Medical Imaging and Graphics 39, 55 –61 (2015), doi:10.1016/j.compmedimag.2014.03.004
18. Koitka, S., Friedrich, C.M.: Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 5-8 2016)
19. Lavrenko, V., Rath, T.M., Manmatha, R.: Holistic word recognition for handwritten historical documents. In: Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on. pp. 278–287 (2004)
20. Müller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF: experimental evaluation in visual information retrieval. Springer-Verlag Berlin Heidelberg (2010), doi:10.1007/978-3-642-15181-1
21. Müller, H., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S.: Creating a classification of image types in the medical literature for visual categorization. In: SPIE Medical Imaging (2012)
22. Pratikakis, I., Zagoris, K., Gatos, B., Louloudis, G., Stamatopoulos, N.: ICFHR 2014 Competition on Handwritten Keyword Spotting (H-KWS 2014). In: Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. pp. 814–819 (Sept 2014)
23. Puigcerver, J., Toselli, A.H., Vidal, E.: Icdar2015 competition on keyword spotting for handwritten documents. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. pp. 1176–1180 (Aug 2015)
24. Ramisa, A., Yan, F., Moreno-Noguer, F., Mikolajczyk, K.: Breakingnews: Article annotation by image and text processing. CoRR abs/1603.07141 (2016), http://arxiv.org/abs/1603.07141
25. Rodríguez-Serrano, J.A., Perronnin, F.: Handwritten word-spotting using hidden Markov models and universal vocabularies. Pattern Recognition 42, 2106–2116 (September 2009)

26. Sorensen, S., Li, P., Kolagunda, A., Jiang, X., Wang, X., Shatkay, H., Kambhamettu, C.: UDEL CIS working notes in ImageCLEF 2016. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 5-8 2016)
27. Sánchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: Icfhr2014 competition on handwritten text recognition on transcriptorium datasets (htrts). In: Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. pp. 785–790 (Sept 2014)
28. Sánchez, J.A., Toselli, A.H., Romero, V., Vidal, E.: Icdar 2015 competition htrts: Handwritten text recognition on the transcriptorium dataset. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. pp. 1166–1170 (Aug 2015)
29. Toselli, A.H., Puigcerver, J., Vidal, E.: Context-aware lattice based filler approach for key word spotting in handwritten documents. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. pp. 736–740 (Aug 2015)
30. Tsikrika, T., de Herrera, A.S., Müller, H.: Assessing the scholarly impact of imageclef. In: Cross Language Evaluation Forum (CLEF 2011). Lecture Notes in Computer Science (LNCS), Springer (2011), doi:10.1007/978-3-642-23708-9_12
31. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., del Mar Roldán García, M.: General Overview of ImageCLEF at the CLEF 2015 Labs. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, vol. 9283, pp. 444–461. Springer International Publishing (2015), doi:10.1007/978-3-319-24027-5_45
32. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2012 Scalable Web Image Annotation Task. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop, Online Working Notes. Rome, Italy (September 17-20 2012), http://ceur-ws.org/Vol-1178/CLEF2012wn-ImageCLEF-VillegasEt2012.pdf
33. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: CLEF2014 Working Notes. CEUR Workshop Proceedings, vol. 1180, pp. 308–328. CEUR-WS.org, Sheffield, UK (September 15-18 2014), http://ceur-ws.org/Vol-1180/CLEF2014wn-Image-VillegasEt2014.pdf
34. Villegas, M., Paredes, R., Thomee, B.: Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain (September 23-26 2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-VillegasEt2013.pdf
35. Villegas, M., Puigcerver, J., Toselli, A.H.: ImageCLEF 2016 Bentham Handwritten Retrieval Dataset (2016), doi:10.5281/zenodo.52994
36. Villegas, M., Puigcerver, J., Toselli, A.H., Sánchez, J.A., Vidal, E.: Overview of the ImageCLEF 2016 Handwritten Scanned Document Retrieval Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 5-8 2016)
37. Wang, J., Gaizauskas, R.: Generating image descriptions with gold standard visual inputs: Motivation, evaluation and baselines. In: Proceedings of the 15th European Workshop on Natural Language Generation (ENLG). pp. 117–126. Association for Computational Linguistics, Brighton, UK (September 2015), http://www.aclweb.org/anthology/W15-4722

38. Zagoris, K., Ergina, K., Papamarkos, N.: Image retrieval systems based on compact shape descriptor and relevance feedback information. Journal of Visual Communication and Image Representation 22(5), 378 – 390 (2011)