2019 15th International Conference on eScience (eScience)

Quality-aware Human-Machine Text Extraction for Biocollections using Ensembles of OCRs

Icaro Alzuru
CISE Department
University of Florida
Gainesville, US
ialzuru@ufl.edu

Rhiannon Stephens
Research Institute
Australian Museum
Sydney, Australia
rhiannon.stephens@austmus.gov.au

Andréa Matsunaga
ACIS Lab.
University of Florida
Gainesville, US
ammatsun@acis.ufl.edu

Maurício Tsugawa
ACIS Lab.
University of Florida
Gainesville, US
tsugawa@ece.ufl.edu

Paul Flemons
Research Institute
Australian Museum
Sydney, Australia
paul.flemons@austmus.gov.au

José A.B. Fortes
ACIS Lab.
University of Florida
Gainesville, US
fortes@acis.ufl.edu

***Abstract*—Information Extraction (IE) from imaged text is affected by the output quality of the text-recognition process. Misspelled or missing text may propagate errors or even preclude IE. Low confidence in automated methods is the reason why some IE projects rely exclusively on human work (crowdsourcing). That is the case of biological collections (biocollections), where the metadata (Darwin-core Terms) found in digitized labels are transcribed by citizen scientists. In this paper, we present an approach to reduce the number of crowdsourcing tasks required to obtain the transcription of the text found in biocollections' images. By using an ensemble of Optical Character Recognition (OCR) engines -- OCRopus, Tesseract, and the Google Cloud OCR -- our approach identifies the lines and characters that have a high probability of being correct. This reduces the need for crowdsourced transcription to be done for only low confidence fragments of text. The number of lines to transcribe is also reduced through hybrid human-machine crowdsourcing where the output of the ensemble of OCRs is used as the first "human" transcription of the redundant crowdsourcing process. Our approach was tested in six biocollections (2,966 images), reducing the number of crowdsourcing tasks by 76% (58% due to lines accepted by the ensemble of OCRs and about 18% due to accelerated convergence when using hybrid crowdsourcing). The automatically extracted text presented a character error rate of 0.001 (0.1%).**

Keywords—OCR, crowdsourcing, biocollections, ensemble, hybrid, information extraction, text extraction, human-machine

I. Introduction

Humans have an extraordinary capacity to extract information from images. For example, from Figure 1, we could say: "*There is a cockroach on a pinned foam. It presumably belongs to the Australian Museum, which assigned code K 482255 to it. The cockroach was captured at the road to Mt. Baldy, in a place with latitude 17.16'S and longitude 145.25'E. The location is at 1,097 meters above the sea level. The cockroach is about 14 mm long and was captured by Rentz and Richardson.*" Besides all this semantic information about what we identified as the main object in the image and the automatic interpretation learned from the text in the labels, we can add characteristics about the labels: "*There are four small labels. From top to bottom, the upper two labels were made in 2011, and they look older than the third one. The second label contains handwritten text. The last label provides the scale and colormap of the photograph.*" The amount of information we can add will depend on our previous knowledge. For example, biologists could probably say more about the cockroach.

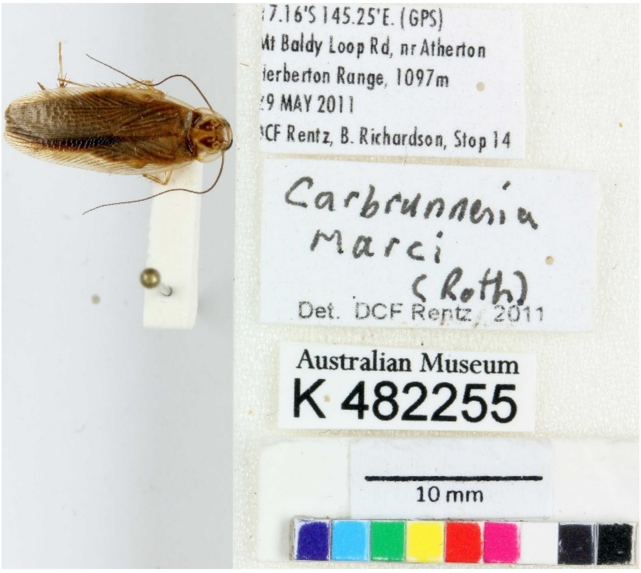


Figure 1. Specimen K 482255 from the Cockroaches Expedition - 2, Australian Museum Entomology collection.

Humans can identify objects in an image, extract knowledge from them, describe objects' characteristics, and even make inferences about what they see, based on previous knowledge.

Artificial intelligence remains unable to analyze an image the way humans do, but it has allowed the creation of algorithms with very specific capabilities that mimic those of humans. With the advent of machine learning, there has been enormous progress on object classification (e.g., type of insect in Figure 1) and optical character recognition (machine-encoding the printed, handwritten, or typed text).

Text extraction -- the identification of each of the characters in an image -- is the main topic of the research reported in this

978-1-7281-2451-3/19/$31.00 ©2019 IEEE
DOI 10.1109/eScience.2019.00-72

paper. It is an fundamental problem because understanding the symbols in the image is a pre-requisite for the extraction of information. If some characters are omitted or misspelled, we may be making false interpretations and introducing or propagating errors about the image and its content.

Standalone optical character recognition (OCR) engines are based on the segmentation of images at the character level and the use of per-symbol individual neural network classifiers. In the last five years, OCR engines have been transformed into text recognition cloud services based on long short-term memory (LSTM) models, with higher character recognition rate and a simplified line-level training [1].

Despite the recent progress in the quality and availability of the OCR technology, general text extraction is still an open problem. Diverse studies [2][3] claim character error rates (CERs) lower than 0.01 (1%) in certain types of documents and fonts. Nevertheless, this cannot be interpreted as the convergence to a final general solution for the text extraction problem.

Two of the most challenging problems for current OCR engines are the segmentation of the image in lines of text and handwritten text recognition. OCRopus [4] and Tesseract [5], arguably the two most popular open source OCR engines, do not provide models to recognize handwritten text. Even the ABBYY FineReader [6], a commercial OCR engine, stops working or generates many misspelling errors when handwritten text is found. At present, to the best of our knowledge, the only OCR engine capable of partially recognizing handwritten text is the Google Cloud OCR (GC-OCR)[7].

TABLE I. OCR ENGINES' OUTPUT COMPARISON

OCRopus 1.3.3	**Tesseract 4.0**	**Google Cloud OCR**
g 92--- ---		**17.16'S 145.25'E.**
. : .-- "-	**Car**h**run**ag,, o	**(GPS)**
C T, --	**Mar**e;	**Mit Baldy Loop Rd, nr**
-+., S ; . A	R, ¢ Le**th)**	**Atherton**
7.l651452SE [6P9]		j**erberton Range,**
Wf **Baldy** loop **d, nr**	**Det. DCF Re**p**t**	**1097m** .
Athen**on**	**2011**	**9 MAY 2011**
ierhen**ton ange**. **1097m**		**CF Rentz, B.**
'9MM**Y 2011**		**Richardson, Stop14**
tCH **Rent**L 3.		**Carbrunnnia**
Rih**ordson, S**!**op 14**		**Marci**
A		s **Roth)**.
s-a t5rend6yi 44		**Det. DCF Rentz 2011**
tri**a rc**		**Australian Museum**
e d 44 Y		**K 482255**
WOMJ		**10 mm**
D**Det. CF R**e**rt**: ed		
Australian Museum		
K 4ocdbo		
g53rw P s		
10 r**mm**		
r-yr--;		
M- Lai ----- --- LuA.--		

Table I shows the output of OCRopus, Tesseract, and the GC-OCR for the image in Figure 1, using their default English recognition models. The characters that match the real text are highlighted in bold. Omitted characters are not represented in the table. We observe many errors in the OCRopus output and almost no output from Tesseract. GC-OCR shows a "close to perfect" output, when compared to the other two engines. OCRopus and Tesseract are highly affected by segmentation (see Section IV.B) and their models are not trained for handwritten text recognition. GC-OCR dynamically uses more than one recognition model.

In the 203 characters of the GC-OCR output, we identify 10 errors, among insertions, omissions, and modifications. This means a CER close to 0.05, which may seem low, but the real problem for IE projects that rely on OCR is confidence. There may be few errors, but what if the errors occur in dates or proper nouns? How will these errors affect posterior processes? Is there any missing word in the extracted text?

This lack of confidence in the text extracted by OCR engines makes IE projects rely on humans to type the information in images. That is the case of projects like Notes from Nature [8] and DigiVol [9], which utilize crowdsourcing (volunteers) for the transcription of Darwin-core (DC) Terms from photographs of specimens in biocollections.

Despite using humans, confidence is key for data that will be used in other scientific studies. Biocollection transcription projects use redundancy to improve the accuracy and increase the confidence in the data. One of the most common approaches is to ask several volunteers to transcribe each image and then use a consensus algorithm to generate the final value (Notes from Nature's approach). Another possibility is to first ask a volunteer to transcribe the DC terms of an image, and then ask a more experienced volunteer to review the transcription (DigiVol's approach).

Progress towards automating IE from biocollections includes the following:

- Sophisticated interfaces to facilitate the load of the information: SALIX [10] loads the results of applying OCR and specific Natural Language Processing (NLP) algorithms into a web form so that users can correct and complete the transcription; ScioTR [11] allows users to select an area of the image and assign the OCR result to a term, optionally editing the value. In these interfaces, users end up searching through the image and reviewing almost every value. Therefore, the IE process is accelerated only when the OCR's output and the NLP algorithms compensate the typing effort of the user.
- The Royal Botanic Garden Edinburgh has accelerated the IE process by using OCR and NLP to automatically extract two terms: Collector and Country [12]. This allows a first classification of the specimen and then for volunteers to complete the transcription of the remaining terms.

Despite these automation attempts, most or all the metadata in biocollections are still extracted by humans. It is difficult to say if a fully automated IE method, with no human review, will become available but an intermediate human-machine solution is certainly possible.

In this paper, we propose a method to reduce the amount of text transcribed by humans, through the automated estimation of confidence in the text extracted by an ensemble of three OCR engines: OCRopus, Tesseract, and GC-OCR. This estimated confidence allows the acceptance of the OCR's output of some segments of text and the decision to request humans to transcribe the remaining text segments.

Furthermore, the crowdsourcing sessions for the transcription of the remaining segments are reduced and accelerated by replacing one of the crowdsourced transcriptions with the output of the ensemble of OCRs.

This paper is about text extraction, which is an intermediate step towards the partial or complete automation of the IE from biocollections. Additional work needs to be done to obtain the terms. We believe that current NLP methods are able to perform the DC Terms extraction if the extracted text is complete and accurate. Moreover, the complexity and training required by the volunteers to transcribe text is smaller than the required to identify and extract DC Terms (domain specific information).

Our ensemble-of-OCRs approach is able to identify as correct 57.55% of the text, with a per-character accuracy of 99.9%. These characters are accepted as the final transcription and are not required to be typed by humans, considering the confidence obtained by using three OCR engines.

Two common crowdsourcing approaches, majority voting and transcriber/reviewer, were modified to use the output of the ensemble of OCRs as the first "human" transcription, for the remaining 42.45% of the text. Using this human-machine approach, the required crowdsourcing tasks were reduced by 37% when using majority voting with three workers and by 50% in the transcriber/reviewer crowdsourcing approach.

In total, the number of crowdsourcing tasks to transcribe the text in the biocollections' images was reduced by 76%.

The code and results of this research are available online at https://github.com/acislab/HuMaIN_Text_Extraction.

II. Related Work

Optical Character Recognition (OCR) technology has recently improved in several ways:

- The font-based neural networks models have been replaced by LSTM networks [2][3][13].
- The OCR desktop applications have been converted to cloud services available everywhere [14][15].
- Handwritten text recognition, previously only available for small corpora, is now included as a model [7][16].

Despite these improvements, there are still challenges in page segmentation, handwritten text recognition, binarization, layout analysis, and post-OCR correction [17]. Most of these challenges show up on historical documents or scene text, where background, graphic elements, and the integrity of the characters (among many other reasons) affect the accuracy of the OCR's output. This unpredictability of the OCR accuracy and its impact on upcoming research are big problems [18].

Several studies have tried to predict the quality or accuracy of the OCR by using a subset of the images [19], latent Dirichlet allocation [20], or Spatial Frequency Response [21], among many other techniques; but these methods provide quality estimations of the entire extracted text. If the confidence in a transcribed document is relatively low, should we discard all the extracted text? Alternatively, if the confidence is high, should we have the same high confidence in every extracted word? It is more useful to predict confidence at a word level.

Some studies improve the character confidence estimation using n-grams [22][23], assigning the probability of a character based on the k previous characters. Our method also uses n-grams, but at a word-level and only to augment the probability, not to correct words. Relying only on n-grams may be risky due to rare n-grams, and the big impact a single character can have on coded or numeric fields such as "year".

To improve robustness, our approach uses an ensemble of OCRs, word-level n-grams, and descriptive statistics at character-level to identify high confidence segments of text.

Previous studies have also used ensembles of OCRs, but they have not used the ensemble with the objective of increasing confidence in the extracted text and identifying correct segments of text. In [24], two OCR engines are used to improve quality, assuming the existence of dictionaries and choosing to evaluate only aligned words. Other researchers generate different versions of the one single image and run the same OCR engine on them to improve the quality of the output [25][26]. Our research uses additional statistical tools to dynamically create the dictionaries. To align the text, we propose a hybrid crowdsourcing mechanism for those segments with low confidence.

Our approach goes beyond improving the quality estimation of the OCR's output. We accept that some problematic cases are going to exist, and that human help is going to be required to extract the final text. Our hybrid human-machine method reduces the amount of data to be transcribed by the crowd.

We use what is called a SELFIE model [27] which is basically a cost-incremental model for the extraction of text; using crowdsourcing as the last self-aware process of the data extraction workflow.

Our hybrid crowdsourcing method was inspired by [28] to minimize the number of crowdsourcing tasks. In the referenced research, human responses and machine classification algorithms are used to identify birds in images. We extended this idea of using the machine results as human results to the area of text extraction.

III. Quality-aware Text Extraction

The objective of this research is to reduce the amount of human work needed to extract text from biocollections' images. We designed a quality-aware approach to decide when to trust and accept automatically extracted characters, words, and lines.

The approach uses the per-character accuracy probability provided by OCR engines and, more importantly, majority voting to increase confidence in the extracted values. The per-character accuracy probabilities of a single OCR engine can be used for output quality estimation, but the confidence in that estimation needs to be increased by other mechanisms, such as redundancy.

In order to trust in the values extracted by OCR engines, we emulate the consensus mechanism applied in crowdsourcing. Three different OCR engines, with different recognition models, are asked to extract the text of the same segments of image (lines). For each line:

- If the outputs of the three OCR engines match, we assume there is a high probability that the value is accurate and the text is accepted. If only two of the outputs match, but the average accuracy probability of both OCR engines for the line is high, we say consensus has been reached and the value is accepted as the final transcription of the line. This line-level agreement corresponds to the first quality-aware process in Figure 2. The intuition behind trusting majority-voted outputs is supported by the probability analysis of this approach, see Section III.A for details.

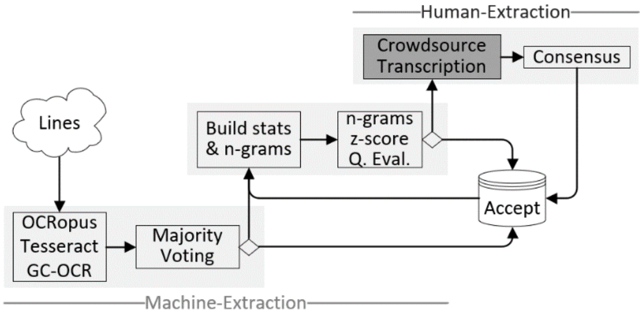


Figure 2. Quality-aware Model for Text Extraction.

- Those lines for which consensus is not reached in the first step are analyzed at a smaller granularity level, trying to build confidence in the characters and words in them. Two types of support data structures are derived:
  - N-grams: 1-gram (unigram) and 2-gram (bigram) models are built from the content in matched lines. N-grams with three or less repetitions are discarded to reduce false positives. Words with less than two characters are not considered for the n-grams.
  - Descriptive statistics: For every possible character, the mean and standard deviation of the OCR engine's accuracy probabilities are computed, using only the accepted lines. A different set of per-character descriptive statistics is computed for each OCR engine.

  A new consensual transcription of every line is built using the outputs of the three OCR engines. The per-character confidence (accuracy probability) is augmented using the n-grams and descriptive character statistics. See the details of this algorithm in Section III.F.

  The consensual transcription is accepted as the final transcription for the line if all its augmented per-character confidence values are equal to 1. This transcription generation and acceptance procedure corresponds to the second quality-aware process in Figure 2.

- The lines for which the transcriptions are not accepted are sent to a crowdsourced processing task (third step in Figure 3). The output of the ensemble of OCRs (the consensual transcriptions) are used as a candidate output in the crowdsourcing tasks to accelerate convergence to a final transcription for the line. Two commonly used crowdsourcing approaches were accelerated using this human-machine collaboration approach. The methods are explained in Section III.B.

The data sets, images and their full text transcription (ground-truth data) utilized in this paper were provided by iDigBio and the Australian Museum. See Section III.C for a detailed description of the six biocollections utilized in the experiments.

These images do not have any specific layout and can contain text of different types (printed, typewritten, handwritten, or stamped), font sizes, colors, backgrounds, and languages (mainly Latin scripts). The text can also be skewed, overlapped with other objects, or underlined. Due to the diversity of the text in these images, the problem of text extraction from biocollections has some similarity to the "*Robust Reading Challenge on Multi-lingual scene text detection and recognition*" of the ICDAR conference [29]. The text in these images seldom contains paragraphs or phrases that follow general grammar rules, but instead, it is typically an unordered set of proper nouns, dates, alphanumerical codes, coordinates, and titles. In this scenario, the use of general dictionaries for error correction may not be an effective alternative.

Three OCR engines are selected for the ensemble. The GC-OCR engine is selected because, to the best of our knowledge, it is the only OCR engine that provides support for general handwritten text recognition. It automatically selects the best recognition model to use in every line of text [30], which is convenient in our case considering the characteristics of the images mentioned before. The GC-OCR is available through the Google Cloud Vision API. It is not free. OCRopus and Tesseract are selected because their recognition models can be extended through training, they are open source, actively improved, and are two of the most commonly used engines.

The machine-only quality baseline, using the out-of-the-box recognition models, was collected for the three OCR engines: OCRopus, Tesseract, and GC-OCR on the images of the data set. For output's quality estimation, the Damerau-Levenshtein similarity is computed (defined in Section III.D) between the OCR engines' result and the ground-truth data. The baseline shows the independent out-of-the-box accuracy of each of the three OCR engines on the data set, see Section IV.A.

Our approach compares the outputs of the OCR engines to generate a new output with augmented confidence. To make this comparison possible, the three engines must work on the same image segments. For this purpose, images were segmented into lines.

Line segmentation is still an open problem [31]. Segmentation errors highly affect the quality of the extraction process and may compromise ideas like the ensemble of OCRs. After testing several methods, including the line segmentation procedures of OCRopus and Tesseract, we decided to adapt the character-level Google Cloud Vision API's output to generate lines' coordinates. See further line segmentation details in Section III.E.

The importance of the line segmentation process is such that after "replacing" the OCRopus' and Tesseract's segmentation procedures, both OCR engines generated higher quality results. See section IV.B.

The OCRopus, Tesseract, and GC-OCR outputs for each of the lines are submitted to the ensemble-of-OCRs method, detailed in Section III.F. The method uses the per-character confidence of the OCR engines, besides generated n-grams and per-character statistics for the collection, to augment the probabilities of the characters and deciding what lines do not need to be crowdsourced due to a high confidence that all the characters in them are correct (these lines are called accepted). The rest of the lines, with uncertainty in one or more of their characters, are crowdsourced. The results obtained after running the ensemble-of-OCRs method on the lines of the data set are shown in Section IV.C.

Algorithm 1 Human-Machine Text Extraction

```
Input: images_dir
Output: labels_dir
for image in images_dir do
    lines ← segment(image)
    for line in lines do
        line_text ← ensemble_ocr(line)
        if accept(line_text) then
            accepted.add(line_text)
        else
            to_crowd.add(line_text)
        end if
    end for
end for
for line_text in to_crowd do
    crowd_out ← crowdsource(get_img(line_text))
end for
labels_dir ← get_labels(accepted, crowd_out)
```

Algorithm 1 offers a simplified high-level view of the dynamic of the entire text extraction process. The `ensemble_ocr` function represents the first two quality-aware text extraction processes in Figure 2.

A. Probability of Error in Ensembles of OCRs.

OCR engines provide an estimation of the confidence or correctness probability for each of the recognized characters. Because of the use of dictionaries for misspelling corrections and syntactic rules, the selected character may not be the character with highest estimated probability. Nevertheless, these engines do not provide an exact explanation of the meaning of these numbers.

For practical purposes, we will assume that the confidence values provided by the OCR engines are the conditional probability of recognizing the character "x" when the value is "x": $P(x|X = x)$, for a specific recognition model.

Given this assumption, the probability of error when an OCR engine has recognized the value as "x" is:

$$P(error) = P(x|X = \bar{x}) = 1 - P(x|X = x).$$

For example, if the OCR engine confidence for a recognized character is $P(x|X = x) = 0.75$, the probability of error will be $P(error) = 1 - 0.75 = 0.25$.

In an ensemble of three OCR engines, with three independent neural network models, if the three engines agree on the same value, an error of the ensemble is only possible if the three engines are wrong:

$$P_{en}(error) = P_1(x|X = \bar{x}) \times P_2(x|X = \bar{x}) \times P_3(x|X = \bar{x})$$

If the three OCR engines report a confidence greater than 0.75, the probability of error for the ensemble will be:

$$P_{en}(error) \leq (1 - 0.75)^3$$

$$P_{en}(error) \leq 0.015625$$

Therefore, with a relatively low confidence of 0.75 in the extracted character, an ensemble of three OCR engines shows a probability of error smaller than 2%, when the three engines agree in the extracted character. This result shows the intuition behind using redundancy to increase confidence.

In the case of using an ensemble of only two OCR engines, for the same confidence of 0.75, the probability of error of the ensemble is

$$P_{en}(error) = 0.0625$$

which is higher than the probability of error for an ensemble of three OCRs, but certainly much lower than 0.25, the initial probability of error for a single OCR engine.

B. Hybrid Human-Machine Crowdsourcing Approaches

In biocollections, and probably in other areas, two common crowdsourcing approaches to agree in a result are:

1) Consensus: several crowdsourcers transcribe the same image and a posterior process, e.g. majority voting, evaluates the values and decides the final result. An odd number of users is usually selected to do the transcription.

2) Hybrid Transcriber/Reviewer: one user makes the initial transcription and an advanced user or expert reviews and validates the transcription. The final transcription is the output of the review.

Both methods generate high confidence because they involve redundancy: several transcriptions of the same image or several people dedicating time to the same image. This redundancy also implies waste of time: transcribing several times the same value or reviewing values that are correct.

Extending the idea applied by Branson et al. [28] for image classification, we propose to include the machine as a member of the crowd and using its output in the generation of the final transcription of the images through crowdsourcing. The two commonly used human-oriented crowdsourcing approaches were adapted as follow:

1) Dynamic Human-Machine Consensus: The ensemble-of-OCRs output is considered as the first transcription. One user transcribes the line, if the result is equal to the ensemble's output, the transcription of the line is accepted and no more crowdsourcing is performed for that line. If the ensemble's output and the user's transcription are different, a second user is asked to transcribe the text in the line. The second user's output is compared with both the ensemble's output and the first user's output. If a match is found, the transcription is accepted. If no match is found, we ask a third and final user to transcribe the text in the line. If, with the output of the third user, there is no match either, the transcription with the highest average DL similarity to the other three values is selected. The results of this process are found in Section IV.D.

2) Hybrid Transcriber/Reviewer: The output of the ensemble of OCRs is considered as the transcription of the non-expert user. A user is asked to review (correct and complete) the transcription. The result of the review process is accepted as the final text of the line. See Section IV.E for the results of applying this crowdsourcing method.

C. Data Set

The images from six biocollections are the data set for the text extraction experiments mentioned in this paper. Three biocollections were prepared by the Augmenting-OCR (A-OCR) Working Group of iDigBio [32], including the entire transcription of text found in the images, made by experts. The other three biocollections belong to DigiVol [9] (The Australian Museum). They include full transcription of the text found in the images, but they were completed by volunteers, therefore containing some omissions and errors. Some characteristics of these biocollections are the following:

- A-OCR Insects (Entomology): 100 images, 33 MB on disk. 1,132 lines of text.
- A-OCR Herbs: 100 images, 124 MB on disk. 3,192 lines of text.
- A-OCR Lichens: 200 images, 31.3 MB on disk. 2,618 lines of text.
- DigiVol Roaches (Cockroaches Expedition-2): 1,117 images, 656 MB on disk. 10,002 lines of text.
- DigiVol Flies (Horse Flies Expedition-3): 1,054 images, 536,6 MB on disk. 7,821 lines of text.
- DigiVol Bees (Carpenter Bees expedition): 395 images, 443 MB on disk. 3,053 lines of text.

D. Damerau-Levenshtein Similarity

The Damerau-Levenshtein distance between two strings is the minimum number of insertions, deletions, substitutions, and transpositions (of adjacent characters) required to convert one string into the other [33].

In order to measure syntactic likeness, we define the Damerau-Levenshtein (DL) similarity between two strings, x and y, as the complement to the normalized DL distance

$$sim_{DL}(x,y) = 1 - \frac{DL\ distance(x,y)}{\max\left(|x|,|y|\right)}$$

where $|x|$ and $|y|$ are the number of characters (size) of strings x and y, respectively.

E. Line Segmentation Approach

The Google Cloud Vision API's output does not provide information about lines of text. It generates coordinates, confidence estimation and values at a page, block, word, and character level, but not at a line level. Internally it does keep track of break lines. Using the coordinates of the individual characters and the registered break lines, we created a program that reconstructed the coordinates of the lines. This method proved to be more accurate than the segmentation performed using the OCRopus script or the Tesseract's hOCR coordinates.

F. Ensemble-of-OCRs Algorithm

The ensemble of OCRs is an extensive algorithm. It receives a set of lines (images) and creates, as output, two directories: a directory with the transcription of the lines with high confidence (accepted lines), and another directory with consensual transcription of the remaining lines (rejected lines).

Algorithm 2 summarizes the steps of this process. The lines are processed in batch, but to facilitate the understanding of the algorithm, it is presented as a line-by-line process.

Algorithm 2 Ensemble of OCRs

```
Input: lines_dir
Output: dir_accepted, dir_to_crowd
for line in lines_dir do:
    ocropus, tesseract, & gc-ocr outputs are collected
    if (two or three outputs match) then:
        accept the common output
Using the output of the lines that matched:
    build n-grams
    build OCRopus' character statistics
    build Tesseract's character statistics
    build GC-OCR's character statistics
for line in non_accepted_lines:
    augment to 1.0 the prob. of words found in n-grams
for line in non_accepted_lines:
    align the three OCR outputs for the line
    for char in line:
       if (statistical consensus reached for char) then:
           augment to 1.0 the probability of the char
    if all the line's chars have probability 1.0 then:
        accept the line
```

OCRopus, Tesseract, and GC-OCR are firstly run on all the lines to generate the confidence probability by character. The outputs of the three OCR engines are compared to each other and the outputs that match are accepted as the final transcription of the correspondent line.

Using the transcription of the lines that matched, the n-grams and the per-character statistics of every OCR engine are built. The remaining lines are scanned, if a word belongs to a n-gram, the probabilities of its characters are made 1.0.

Then, the outputs of the OCR engines for each line are aligned and a per-character evaluation is performed to construct a new line transcription. If for a given position of the alignment the three engines extracted the same character, then this character is accepted with a probability of 1.0. If only two OCR engines agree in the value, their z-score for the character accuracy probability is computed; if both z-score are greater than 0.5, the character is accepted and its probability is also made 1.0.

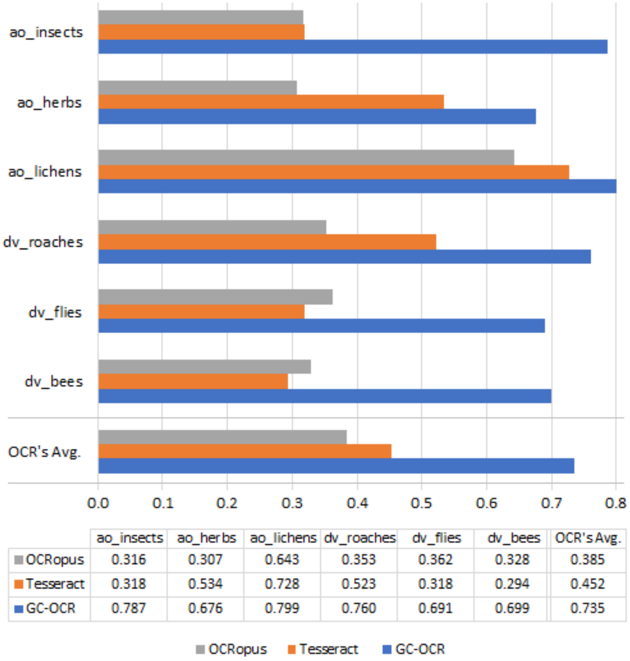
If consensus is not reached for a certain position of the alignment, the character extracted by the OCR engine with the highest general accuracy is selected; in our experiments the GC-OCR is the highest-accuracy OCR engine, see Section IV.B. After evaluating all the lines at a character level, the lines with an average character accuracy of 1.0 are accepted; i.e., lines are accepted if all their characters belong to words in n-grams or consensus at character level was reached.

IV. Results

This section shows the numerical results obtained for the Quality-aware Text Extraction approach and the steps presented in Section III.

A. Baseline – Out-of-the-box OCR Engines' Accuracy

OCRopus, Tesseract, and GC-OCR were independently run on the images of the data set. OCRopus and Tesseract were run using their respective out-of-the-box English recognition models. The extracted text was compared to the human transcription (ground-truth data) using the Damerau-Levenshtein similarity metric. Figure 3 shows the obtained similarity, per OCR engine and per biocollection.



	ao_insects	ao_herbs	ao_lichens	dv_roaches	dv_flies	dv_bees	OCR's Avg.
▨ OCRopus	0.316	0.307	0.643	0.353	0.362	0.328	0.385
▨ Tesseract	0.318	0.534	0.728	0.523	0.318	0.294	0.452
▨ GC-OCR	0.787	0.676	0.799	0.760	0.691	0.699	0.735

Figure 3. Damerau-Levenshtein similarity of the OCRopus, Tesseract, and GC-OCR outputs to the ground-truth data, per biocollection. Range: 0.0 to 1.0. A similarity of 1.0 corresponds to the case of two identical strings.

The quality of the output generated by the GC_OCR engine is higher than the output quality of OCRopus and Tesseract in every biocollection. The items in the ao_lichens biocollection are images of text, i.e. the images do not contain the specimen, rulers, or other objects; this was the biocollection where the three OCR engines reached their highest quality.

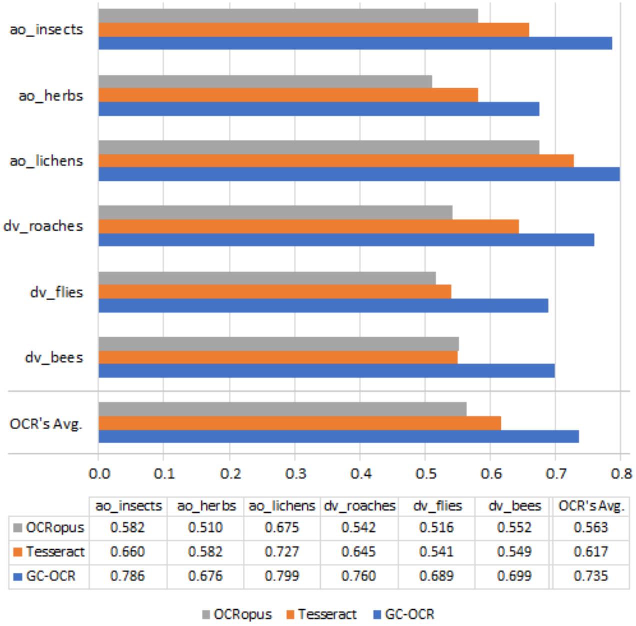
GC-OCR, which obtained the highest average quality, did not get an average similarity greater than 0.8 in any biocollection. Its average global similarity is 0.735, which still makes it difficult to implement an IE process that depends on this text.

B. OCR Engines' Accuracy After Line Segmentation

In the results shown in Figure 3, each OCR engine uses its own binarization and segmentation algorithm, i.e. each engine segments each image in a different way. We need a common or standard segmentation to be able to compare the outputs of the OCR engines. This common segmentation method was explained in Section III.E. OCRopus' and Tesseract's segmentation functions are not used in the ensemble-of-OCRs approach.

Figure 4 shows the DL similarity to the ground-truth data of OCRopus, Tesseract, and GC-OCR using the line segmentation procedure explained in Section III.E. As expected, the GC-OCR output quality is basically the same as in Figure 3, with small differences probably generated by the real line segmentation internally utilized by the GC-OCR. On the other hand, OCRopus and Tesseract improved their average output quality in 46% and 37% compared to Figure 3, where they used their own segmentation algorithm. This result shows the importance of the segmentation process for the OCR's output quality.

The per-engine transcription generated for each line are the input for the ensemble-of-OCRs process described in IV.C.



	ao_insects	ao_herbs	ao_lichens	dv_roaches	dv_flies	dv_bees	OCR's Avg.
▨ OCRopus	0.582	0.510	0.675	0.542	0.516	0.552	0.563
▨ Tesseract	0.660	0.582	0.727	0.645	0.541	0.549	0.617
▨ GC-OCR	0.786	0.676	0.799	0.760	0.689	0.699	0.735

Figure 4. Damerau-Levenshtein similarity of the OCRopus, Tesseract, and GC-OCR outputs to the ground-truth data, per biocollection. Range: 0.0 to 1.0. A similarity of 1.0 corresponds to the case of two identical strings.

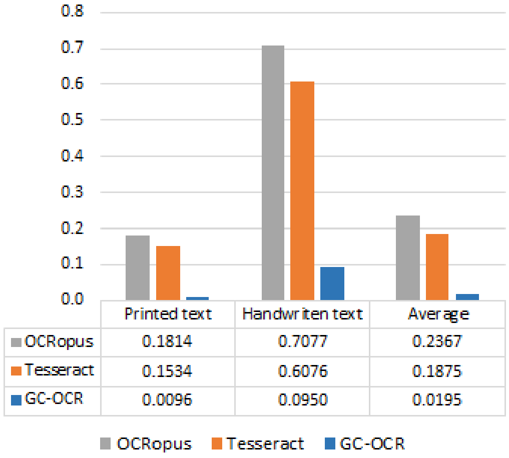
We estimated the character error rate (CER) of the OCR engines' recognition models. For this purpose, a subset of 60 images was analyzed with 10 images randomly selected from each biocollection. Their lines were transcribed to generate their ground-truth data and making possible a per-character evaluation. The characteristics of this subset are detailed in Table II.

TABLE II. Composition of the Subset of 60 Images

Collection	# of Images	# of Lines	Printed Text	Handwritten Text (HRT)	No Text
ao_insects	10	112	104	8	0
ao_herbs	10	320	276	38	6
ao_lichens	10	128	125	3	0
dv_roaches	10	89	83	6	0
dv_flies	10	72	50	19	3
dv_bees	10	80	61	17	2
Total	60	801	699	91	11

Figure 5 shows the CER of OCRopus, Tesseract, and GC-OCR when they are run on the lines of the 60 images subset. OCRopus and Tesseract recognize less than 30% and 40% of the handwritten text characters, respectively. This result is understandable because their models were not trained for this type of text. However, their CER for printed text was also high, considering the ultimate goal of extracting DC terms and using them in posterior scientific studies.

The GC-OCR's CER is less than 0.02, on average, and less than 0.01 for printed text, which includes typewritten, printed and stamped text. This is the error rate in the cropped lines. The last column of Table II shows the number of cropped lines that do not really contain any text, i.e. segmentation errors generated by the erroneous identification of text. Moreover, part of the text was not included in the cropped lines. Therefore, we can conclude that the GC-OCR's CER is at least 0.02.



	Printed text	Handwr iten text	Average
■ OCRopus	0.1814	0.7077	0.2367
■ Tesseract	0.1534	0.6076	0.1875
■ GC-OCR	0.0096	0.0950	0.0195

Figure 5. Character Error Rate (CER) of the OCRopus, Tesseract, and GC-OCR engines for the subset of 60 images described in Table II.

The problem of hypothetically using only the GC-OCR to extract the metadata from the biocollections images is that we do not know the handwritten composition of the images, we cannot predict when segmentation is not accurate, and we do not know where the errors are. Moreover, if we run NLP processes on the resulting text, we may be propagating errors and introducing inaccuracy in the extracted information and the applications that will posteriorly use it. All this generates uncertainty in the extracted text.

To increase confidence in the extracted data, we use several OCR engines (an ensemble) to detect those segments of text where two or more OCR engines agree in their content. This increases trust in the automatically extracted data and reduces the amount of work assigned to humans.

C. *Ensemble of OCRs*

After segmenting the images in lines, the lines are processed by the ensemble-of-OCRs algorithm explained in Section III.F. The result of this algorithm is a set of accepted transcriptions (lines for which we are confident that their automated transcription is correct) and a set of images and their transcriptions which will be sent to be crowdsourced because their confidence is not high enough to be accepted. The performance of the ensemble of OCRs for each of the biocollections of our data set is shown in Table III.

TABLE III. PERCENTAGE OF ACCEPTED LINES PER BIOCOLLECTION

	Images	Lines	Accept	Crowd	% Accepted
ao_insects	100	1132	711	421	62.81%
ao_herbs	100	3192	1657	1535	51.91%
ao_lichens	200	2618	1639	979	62.61%
dv_roaches	1117	10002	5831	4171	58.30%
dv_flies	1054	7821	4372	3449	55.90%
dv_bees	395	3053	1800	1253	58.96%

In total, 57.55% (16,010) of the 27,818 lines were accepted using the ensemble-of-OCRs algorithm. The algorithm uses majority voting, n-grams, and descriptive statistics to decide when a transcription must be accepted. From each biocollection, 100 accepted lines were randomly selected, making a total of 600 lines, to be reviewed by crowdsourcing participants.

Of the 10,081 characters in the 600 lines, the users made changes, insertions, or deletions in only 10 characters. This means that the accepted lines have a CER of 0.001 and an accuracy of 99.9%. This CER is better than the average CER of 0.0195 obtained by GC-OCR for a subset of lines, see Figure 5.

D. *Crowdsourcing - Dynamic Human-Machine Consensus*

For experimental purposes, 100 lines per collection, 600 in total, were randomly selected from the 11,808 lines that were not accepted in the ensemble-of-OCRs process. Users were asked to transcribe the content of the 600 lines. They did not know the content of the transcription generated using the ensemble of OCRs. A single user per line completed the transcription.

These crowdsourced transcriptions were compared to the ensemble-of-OCRs output. There was a match in 230 (38.3%) of the 600 lines. Their content was accepted. For the remaining 370 lines, a second crowdsourcing round was requested to the volunteers of DigiVol.

After processing the transcribed data, there was a match in 34 lines (5.67%) between the ensemble of OCRs and the transcriptions of the second round. There was a match in 176 lines (29.3%) between the transcriptions of the first and second crowdsourcing rounds.

In total, after the two crowdsourcing rounds, the transcriptions of 160 lines (26.67%) did not reach consensus with the ensemble-of-OCRs output or between them. A last third round of crowdsourced transcription was done for these remaining 160 images.

After comparing the result of the third crowdsourcing round to the outputs of the ensemble of OCRs, the first, and the second rounds of crowdsourced transcriptions, a match was found in 9, 36, and 38 lines, respectively; totaling 83 lines (13.83%) and leaving 77 lines (12.83%) where a match was not found. In these lines, other mechanisms to reach consensus can be applied, but those methods are not the focus of this paper.

In summary, 38.3% of the 600 lines required a single human transcription, 35.0% required two human transcriptions, and 13.83% required three human transcriptions. After applying this dynamic consensus approach, in 12.83% of the lines, neither hybrid nor human consensus could be reached using three human transcriptions.

Assume that nL represents the number of lines to transcribe through dynamic human-machine consensus. Assuming the proportion of matches found in the experiments above generalize, we can state the following:

- The original, human-only approach requires $3 \times nL$ crowdsourcing tasks (human transcriptions).
- 1-transcription matches require $0.3833 \times nL$ crowdsourcing tasks.
- 2-transcription matches require $0.7 \times nL$ tasks.
- 3-transcription matches require $0.8 \times nL$ tasks.

Therefore, using dynamic human-machine consensus, we save about $\frac{3 - 0.3833 - 0.7 - 0.8}{3} => 37.22\%$ of the crowdsourcing tasks, when compared to the human-only version of majority voting in crowdsourcing.

E. *Crowdsourcing – Hybrid Transcriber/Reviewer*

The Australian Museum with the DigiVol platform uses another method to find the final transcriptions. One user transcribes the text and then an advanced user reviews the transcription, the result of this review being accepted as the final transcription.

In our hybrid human-machine approach, the output of the ensemble of OCRs (which did not get confidence high enough to be accepted), is sent to the reviewer.

For this crowdsourcing method a different subset of 100 lines per collection, 600 in total, were randomly selected from lines not accepted by the ensemble of OCRs. Users were asked to review (correct or complete) the content of the 600 lines. One single user completed the review of any given line.

In the review process of the 9,025 characters in the 600 lines, the reviewers detected 356 misspelled characters, 288 omissions, and 89 non-existent characters, for a total CER of 0.081. In terms of human effort, the review process was completely done, arguably with equivalent effort to what would require if humans had completed the first transcription, but the entire initial human transcription was saved. Therefore, this human-machine crowdsourcing approach saved 50% of the transcription tasks compared to its human-only version.

F. *Total Savings in the Number of Crowdsourcing Tasks*

If we assume that the behavior presented in IV.D and IV.E for hybrid crowdsourcing will be similar for the rest of lines for which consensus was not reached in the ensemble-of-OCRs method, the final number of crowdsourcing tasks that were saved by the quality-aware human machine text extraction would be as shown in Table IV.

TABLE IV. SAVINGS IN THE NUMBER OF CROWDSOURCING TASKS

	Tasks required	**Ensemble savings**	**Hybrid crowd. savings**	**Total savings**
Dynamic Human-Machine Consensus	3 x nL	57.55%	15.801%	73.35%
Hybrid Transcriber /Reviewer	2 x nL	57.55%	21.225%	78.78%

In summary, the quality-aware human-machine text extraction approach, using an ensemble of OCRs, saves about 76% of the crowdsourcing tasks when compared to current human-only text extraction approaches.

Conclusions

Despite continuing improvements, OCR engines still make mistakes when automatically extracting the text from the challenging images found in biocollections. This drives information extraction projects to rely exclusively on the transcriptions of citizen scientists.

Given the difficulty of the text-extraction task in these types of images, humans (crowdsourcing) are needed for some especially challenging segments of text. But for the rest of the text, the transcription automatically generated by the OCR engines can be accepted as correct.

In this research, we proposed and showed how to identify segments of automatically extracted text that are correct, by using an ensemble of three OCR engines: OCRopus, Tesseract, and the Google Cloud OCR. Using their outputs, the associated per-character probability and several statistical methods, we were able to detect when the text is correct with an accuracy of 99.9%.

For the biocollections used in our studies, 58% of the automatically extracted text could be identified as correct. The rest was sent to humans for further processing. The consensual transcription of the ensemble of OCRs was used to reduce the number of crowdsourcing tasks.

Two common approaches for the generation of the final transcription in crowdsourcing experiments were tested. On average, the use of the ensemble-of-OCRs result reduced the crowdsourcing tasks by 44%.

In total, considering the reduction in crowdsourcing tasks due to accepting part of the text automatically extracted by the ensemble of OCRs, and the reduction in crowdsourcing tasks by using the ensemble of OCRs as a member of the crowd, our text extraction approach reduced the number of crowdsourcing tasks by 76%.

This result suggests that our approach can be applied to more efficiently use the time of the citizen scientists and to accelerate text and information extraction projects.

Further research is required to further improve the segmentation of text lines, which proved to be crucial for the OCR process.

Acknowledgment

This work is supported in part by the National Science Foundation (NSF) grant No. ACI-1535086 and No. EF-1115210, and the AT&T Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or the AT&T Foundation.

We thank the volunteers of the Australian Museum and DigiVol who participated in the crowdsourcing experiments. Your help made this research possible.

References

[1] C. Reul, U. Springmann, C. Wick, and F. Puppe, “State of the Art Optical Character Recognition of 19th Century Fraktur Scripts using Open Source Engines,” arXiv:1810.03436 [cs], Oct. 2018.

[2] A. Ul-Hasan, F. Shafait, and T. Breuel, “High-Performance OCR for Printed English and Fraktur using LSTM Networks,” presented at the Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2013.

[3] T. M. Breuel, “High Performance Text Recognition Using a Hybrid Convolutional-LSTM Implementation,” in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, vol. 01, pp. 11–16.

[4] T. M. Breuel, “The OCRopus open source OCR system,” in Document Recognition and Retrieval XV, 2008, vol. 6815, p. 68150F.

[5] R. Smith, “An Overview of the Tesseract OCR Engine,” in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, vol. 2, pp. 629–633.

[6] “ABBYY FineReader 14.” [Online]. Available: https://www.abbyy.com/en-us/finereader/. [Accessed: 20-May-2019].

[7] “Detect Handwriting (OCR), Cloud Vision API,” Google Cloud. [Online]. Available: https://cloud.google.com/vision/docs/handwriting. [Accessed: 20-May-2019].

[8] “Notes from Nature.” [Online]. Available: https://www.zooniverse.org/organizations/md68135/notes-from-nature. [Accessed: 20-May-2019].

[9] “DIGIVOL” [Online]. Available: https://digivol.ala.org.au/. [Accessed: 20-May-2019].

[10] A. Barber, D. Lafferty, and L. Landrum, “The SALIX Method: A semi-automated workflow for herbarium specimen digitization,” Taxon, vol. 62, pp. 581–590, Jun. 2013.

[11] “ScioChronicle: ScioTR Available Now in the Win8 Store!,” ScioChronicle, 15-May-2014 . [Online]. Available: http://sciochronicle.blogspot.com/. [Accessed: 20-May-2019].

[12] R. E. Drinkwater, R. W. N. Cubey, and E. M. Haston, “The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels,” PhytoKeys, no. 38, pp. 15–30, May 2014.

[13] “Tesseract version 4.0 releases with new LSTM based engine, and an updated build system,” Packt Hub, 30-Oct-2018. [Online]. Available: https://hub.packtpub.com/tesseract-version-4-0-releases-with-new-lstm-based-engine-and-an-updated-build-system/. [Accessed: 20-May-2019].

[14] J. Walker, Y. Fujii, A. C. Popat, "A web-based ocr service for documents", Proceedings of the 13th International Workshop on Document Analysis Systems. IEEE, Apr. 2018.

[15] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig, “OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym,” in Advances in Visual Computing, 2016, pp. 735–746.

[16] R. R. Ingle, Y. Fujii, T. Deselaers, J. Baccash, and A. C. Popat, “A Scalable Handwritten Text Recognition System,” arXiv:1904.09150 [cs], Apr. 2019.

[17] “Competitions – ICDAR2019.” [Online]. Available: https://icdar2019.org/competitions-2/. [Accessed: 20-May-2019].

[18] M. C. Traub, J. van Ossenbruggen, and L. Hardman, “Impact Analysis of OCR Quality on Research Tasks in Digital Archives,” in Research and Advanced Technology for Digital Libraries, 2015, pp. 252–263.

[19] C. Clausner, S. Pletschacher, and A. Antonacopoulos, “Quality Prediction System for Large-Scale Digitisation Workflows,” in 2016 12th IAPR Workshop on Document Analysis Systems (DAS), 2016, pp. 138–143.

[20] X. Peng, H. Cao, and P. Natarajan, “Document image OCR accuracy prediction via latent Dirichlet allocation,” in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 771–775.

[21] P. K. Rai, S. Maheshwari, and V. Gandhi, “Document Quality Estimation Using Spatial Frequency Response,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1233–1237.

[22] A. Poznanski and L. Wolf, “CNN-N-Gram for Handwriting Word Recognition,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2305–2314.

[23] S. K. Ghosh, E. Valveny, and A. D. Bagdanov, “Visual Attention Models for Scene Text Recognition,” in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, vol. 01, pp. 943–948.

[24] Z. Shan and H. Cao, “Ensemble Optical Character Recognition Systems via Machine Learning,” Course project, 2013. p. 4. [Online]. Available: http://www.zifeishan.org/files/ensemble-ocr.pdf/. [Accessed: 20-May-2019].

[25] I. Q. Habeeb, Z. Q. Al-Zaydi, and H. N. Abdulkhudhur, “Enhanced Ensemble Technique for Optical Character Recognition,” in New Trends in Information and Communications Technology Applications, 2018, pp. 213–225.

[26] W. B. Lund, D. J. Kennard, and E. K. Ringger, “Why Multiple Document Image Binarizations Improve OCR,” in Proceedings of the 2Nd International Workshop on Historical Document Imaging and Processing, New York, NY, USA, 2013, pp. 86–93.

[27] I. Alzuru, A. Matsunaga, M. Tsugawa, and J. A. B. Fortes, “SELFIE: Self-Aware Information Extraction from Digitized Biocollections,” in 2017 IEEE 13th International Conference on e-Science (e-Science), 2017, pp. 69–78.

[28] S. Branson, G. Van Horn, C. Wah, P. Perona, and S. Belongie, “The Ignorant Led by the Blind: A Hybrid Human–Machine Vision System for Fine-Grained Categorization,” Int J Comput Vis, vol. 108, no. 1, pp. 3–29, May 2014.

[29] “ICDAR2019 Robust Reading Competition – Challenge on Multi-lingual Scene Text Detection and Recognition.” [Online]. Available: http://rrc.cvc.uab.es/files/RRC-MLT-2019-CFP.pdf. [Accessed: 20-May-2019].

[30] Y. Fujii, “Optical Character Recognition Research at Google,” in 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), 2018, pp. 265–266.

[31] T. Grüning, R. Labahn, M. Diem, F. Kleber, and S. Fiel, “READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents,” arXiv:1705.03311 [cs], May 2017.

[32] “iDigBio Augmenting OCR Working Group & Hackathon,” GitHub. [Online]. Available: https://github.com/idigbio-aocr. [Accessed: 20-May-2019].

[33] W. H. Gomaa and A. A. Fahmy, “A Survey of Text Similarity Approaches,” International Journal of Computer Applications (0975 – 8887), Volume 68– No.13, April 2013.